Journal of Machine Learning Research 11 (2010) 789-814 Submitted 11/08; Revised 1/10; Published 2/10

Stability Bounds for Stationary φ-mixing and β-mixing Processes

Mehryar Mohri MOHRI@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences
and Google Research
251 Mercer Street
New York, NY 10012

Afshin Rostamizadeh ROSTAMI@CS.NYU.EDU
Department of Computer Science
Courant Institute of Mathematical Sciences
251 Mercer Street
New York, NY 10012

Editor: Ralf Herbrich

Abstract

Most generalization bounds in learning theory are based on some measure of the complexity of the hypothesis class used, independently of any algorithm. In contrast, the notion of algorithmic stability can be used to derive tight generalization bounds that are tailored to specific learning algorithms by exploiting their particular properties. However, as in much of learning theory, existing stability analyses and bounds apply only in the scenario where the samples are independently and identically distributed. In many machine learning applications, however, this assumption does not hold. The observations received by the learning algorithm often have some inherent temporal dependence.

This paper studies the scenario where the observations are drawn from a stationary φ-mixing or β-mixing sequence, a widely adopted assumption in the study of non-i.i.d. processes that implies a dependence between observations weakening over time. We prove novel and distinct stability-based generalization bounds for stationary φ-mixing and β-mixing sequences. These bounds strictly generalize the bounds given in the i.i.d. case and apply to all stable learning algorithms, thereby extending the use of stability-bounds to non-i.i.d. scenarios.

We also illustrate the application of our φ-mixing generalization bounds to general classes of learning algorithms, including Support Vector Regression, Kernel Ridge Regression, and Support Vector Machines, and many other kernel regularization-based and relative entropy-based regularization algorithms. These novel bounds can thus be viewed as the first theoretical basis for the use of these algorithms in non-i.i.d. scenarios.

Keywords: learning in non-i.i.d. scenarios, weakly dependent observations, mixing distributions, algorithmic stability, generalization bounds, learning theory

1. Introduction

Most generalization bounds in learning theory are based on some measure of the complexity of the hypothesis class used, such as the VC-dimension, covering numbers, or Rademacher complexity. These measures characterize a class of hypotheses, independently of any algorithm. In contrast, the notion of algorithmic stability can be used to derive bounds that are tailored to specific learning algorithms and exploit their particular properties. A learning algorithm is stable if the hypothesis it outputs varies in a limited way in response to small changes made to the training set. Algorithmic stability has been used effectively in the past to derive tight generalization bounds (Bousquet and Elisseeff, 2001, 2002; Kearns and Ron, 1997).

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, this assumption, however, does not hold; in fact, the i.i.d. assumption is not tested or derived from any data analysis. The observations received by the learning algorithm often have some inherent temporal dependence. This is clear in system diagnosis or time series prediction problems. Clearly, prices of different stocks on the same day, or of the same stock on different days, may be dependent. But, a less apparent time dependency may affect data sampled in many other tasks as well.

This paper studies the scenario where the observations are drawn from a stationary φ-mixing or β-mixing sequence, a widely adopted assumption in the study of non-i.i.d. processes that implies a dependence between observations weakening over time (Yu, 1994; Meir, 2000; Vidyasagar, 2003; Lozano et al., 2006; Mohri and Rostamizadeh, 2007). We prove novel and distinct stability-based generalization bounds for stationary φ-mixing and β-mixing sequences. 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. Our proofs are based on the independent block technique described by Yu (1994) and attributed to Bernstein (1927), which is commonly used in such contexts. However, our analysis somewhat differs from previous uses of this technique in that the blocks of points we consider are not necessarily of equal size.

For our analysis of stationary φ-mixing sequences, we make use of a generalized version of McDiarmid's inequality given by Kontorovich and Ramanan (2008) that holds for φ-mixing sequences. This leads to stability-based generalization bounds with the standard exponential form. Our generalization bounds for stationary β-mixing sequences cover a more general non-i.i.d. scenario and use the standard McDiarmid's inequality, however, unlike the φ-mixing case, the β-mixing bound presented here is not a purely exponential bound and contains an additive term depending on the mixing coefficient.

We also illustrate the application of our φ-mixing generalization bounds to general classes of learning algorithms, including Support Vector Regression (SVR) (Vapnik, 1998), Kernel Ridge Regression (Saunders et al., 1998), and Support Vector Machines (SVMs) (Cortes and Vapnik, 1995). Algorithms such as SVR 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 (Müller et al., 1997; Mattera and Haykin, 1999). However, to our knowledge, the use of these algorithms in non-i.i.d. scenarios has not been previously supported by any theoretical analysis. The stability bounds we give for SVR, SVMs, and many other kernel regularization-based and relative entropy-based regularization algorithms can thus be viewed as the first theoretical basis for their use in such scenarios.

The following sections are organized as follows. In Section 2, we introduce the definitions relevant to the non-i.i.d. problems that we are considering and discuss the learning scenarios in that context. Section 3 gives our main generalization bounds for stationary φ-mixing sequences based on stability, as well as the illustration of its applications to general kernel regularization-based algorithms, including SVR, KRR, and SVMs, as well as to relative entropy-based regularization algorithms. Finally, Section 4 presents the first known stability bounds for the more general stationary β-mixing scenario.

2. Preliminaries

We first introduce some standard definitions for dependent observations in mixing theory (Doukhan, 1994) 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 $\Pr[Z_k \mid Z_i]$, that is, conditional probabilities may vary at different points in time. 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 these quantities, we are adopting here a version convenient for our analysis, as in Yu (1994).

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*,* $i \le k \le j$*. Then, for any positive integer* k*, the* β*-mixing and* φ*-mixing coefficients of the stochastic process* $\mathbf{Z}$ *are defined as*

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

$\mathbf{Z}$ *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* $\beta(k) \le \beta_0 / k^r$ *(resp.* $\varphi(k) \le \varphi_0 / k^r$*) for all* k*,* exponentially mixing *if there exist real numbers* β_0 *(resp.* $\varphi_0 > 0$*),* β_1 *(resp.* $\varphi_1 > 0$*) and* $r > 0$ *such that* $\beta(k) \le \beta_0 \exp(-\beta_1 k^r)$ *(resp.* $\varphi(k) \le \varphi_0 \exp(-\varphi_1 k^r)$*) for all* k*.*

Both $\beta(k)$ and $\varphi(k)$ measure the dependence of an event on those that occurred more than k units of time in the past. β-mixing is a weaker assumption than φ-mixing and thus covers a more general non-i.i.d. scenario.

This paper gives stability-based generalization bounds both in the φ-mixing and β-mixing case. The β-mixing bounds cover a more general case of course, however, the φ-mixing bounds are simpler and admit the standard exponential form. The φ-mixing bounds are based on a concentration inequality that applies to φ-mixing processes only. Except for the use of this concentration bound and two lemmas 5 and 6, all of the intermediate proofs and results to derive a φ-mixing bound in Section 3 are given in the more general case of β-mixing sequences.

It has been argued by Vidyasagar (2003) that β-mixing is "just the right" assumption for the analysis of weakly-dependent sample points in machine learning, in particular because several PAC-learning results then carry over to the non-i.i.d. case. Our β-mixing generalization bounds further contribute to the analysis of this scenario.[1]

1. Some results have also been obtained in the more general context of α-mixing but they seem to require the stronger condition of exponential mixing (Modha and Masry, 1998).

We describe in several instances the application of our bounds in the case of algebraic mixing. Algebraic mixing is a standard assumption for mixing coefficients that has been adopted in previous studies of learning in the presence of dependent observations (Yu, 1994; Meir, 2000; Vidyasagar, 2003; Lozano et al., 2006). Let us also point out that mixing assumptions can be checked in some cases such as with Gaussian or Markov processes (Meir, 2000) and that mixing parameters can also be estimated in such cases.

Most previous studies use a technique originally introduced by Bernstein (1927) based on *independent blocks* of equal size (Yu, 1994; Meir, 2000; Lozano et al., 2006). 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 Yu (1994).

Lemma 3 (Yu, 1994, 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,*

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

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 ($Y \subseteq \mathbb{R}$ 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 $c \colon Y \times Y \to \mathbb{R}_+$. 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 often 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 $S = (z_1, \ldots, z_m)$:

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

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 $\mathbf{Z}$ 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:

$$R(h_S) = \mathop{\mathrm{E}}_{z}[c(h_S, z) \mid S]. \tag{1}$$

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:

$$R(h_S) = \mathop{\mathrm{E}}_{z}[c(h_S, z) \mid S] = \mathop{\mathrm{E}}_{z}[c(h_S, z)].$$

This setting seems less natural since, if samples are dependent, 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 Yu (1994), Meir (2000), Vidyasagar (2003) and Lozano et al. (2006), even when examining specifically a time series prediction problem (Meir, 2000). Thus, the bounds derived in these studies cannot be directly applied to the more general setting. We will consider instead the most general setting with the definition of the generalization error based on Equation 1. Clearly, our analysis also applies to the less general setting just discussed as well.

Let us also briefly discuss the more general scenario of *non-stationary* mixing sequences, that is one where the distribution may change over time. Within that general case, the generalization error of a hypothesis h_S, defined straightforwardly by

$$R(h_S, t) = \mathop{\mathrm{E}}_{z_t \sim \sigma_t^t}[c(h_S, z_t) \mid S],$$

would depend on the time t and it may be the case that $R(h_S, t) \neq R(h_S, t')$ for $t \neq t'$, making the definition of the generalization error a more subtle issue. To remove the dependence on time, one could define a weaker notion of the generalization error based on an expected loss over all time:

$$R(h_S) = \mathop{\mathrm{E}}_{t}[R(h_S, t)].$$

It is not clear however whether this term could be easily computed and be useful. A stronger condition would be to minimize the generalization error for any particular target time. Studies of this type have been conducted for smoothly changing distributions, such as in Zhou et al. (2008), however, to the best of our knowledge, the scenario of a both non-identical and non-independent sequences has not yet been studied.

3. φ-Mixing Generalization Bounds and Applications

This section gives generalization bounds for $\widehat{\beta}$-stable algorithms over a mixing stationary distribution.[2] The first two sections present our supporting lemmas which hold for either β-mixing or φ-mixing stationary distributions. In the third section, we will briefly discuss concentration inequalities that apply to φ-mixing processes only. Then, in the final section, we will present our main results.

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

The condition of $\widehat{\beta}$-stability is an algorithm-dependent property first introduced by Devroye and Wagner (1979) and Kearns and Ron (1997). It has been later used successfully by Bousquet and Elisseeff (2001, 2002) 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 cause large deviations in its output. The following gives the precise technical definition.

Definition 4 *A learning algorithm is said to be (uniformly)* $\widehat{\beta}$-stable *if the hypotheses it returns for any two training samples S and S' that differ by removing a single point satisfy*

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

We note that a $\widehat{\beta}$-stable algorithm is also stable with respect to *replacing* a single point. Let S and S_i be two sequences differing in the ith coordinate, and $S_{/i}$ be equivalent to S and S_i but with the ith point removed. Then for a $\widehat{\beta}$-stable algorithm we have,

$$\begin{aligned} |c(h_S, z) - c(h_{S_i}, z)| &= |c(h_S, z) - c(h_{S_{/i}}) + c(h_{S_{/i}}) - c(h_{S_i}, z)| \\ &\leq |c(h_S, z) - c(h_{S_{/i}})| + |c(h_{S_{/i}}) - c(h_{S_i}, z)| \\ &\leq 2\widehat{\beta}. \end{aligned}$$

The use of stability in conjunction with McDiarmid's inequality will allow us to derive generalization bounds. McDiarmid's inequality is an exponential concentration bound of the form

$$\Pr[|\Phi - \mathrm{E}[\Phi]| \geq \epsilon] \leq \exp\left(-\frac{m\epsilon^2}{\tau^2}\right),$$

where the probability is over a sample of size m and where $\frac{\tau}{m}$ is the Lipschitz parameter of Φ, with τ a function of m. Unfortunately, this inequality cannot be applied when the sample points are not distributed in an i.i.d. fashion. We will use instead a result of Kontorovich and Ramanan (2008) that extends McDiarmid's inequality to φ-mixing distributions (Theorem 8). 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.

Let us first take a brief look at the problem faced when attempting to give stability bounds for dependent sequences and give some idea of our solution for that problem. The stability proofs given by Bousquet and Elisseeff (2001) assume the i.i.d. property, thus replacing an element in a sequence with another does not affect the expected value of a random variable defined over that sequence. In other words, the following equality holds,

$$\mathop{\mathrm{E}}_{S}[V(Z_1, \ldots, Z_i, \ldots, Z_m)] = \mathop{\mathrm{E}}_{S, Z'}[V(Z_1, \ldots, Z', \ldots, Z_m)], \tag{2}$$

for a random variable V that is a function of the sequence of random variables $S = (Z_1, \ldots, Z_m)$. However, clearly, if the points in that sequence S are dependent, this equality may not hold anymore.

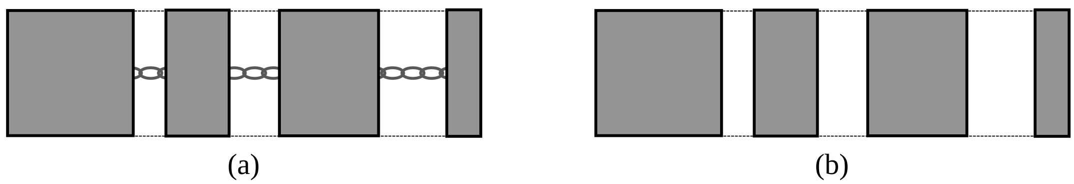

Figure 1: Illustration of dependent (a) and independent (b) blocks. Although there is no dependence between blocks of points in (b), the distribution within each block remains the same as in (a) and thus points within a block remain dependent.

The main technique to cope with this problem is based on the so-called "independent block sequence" originally introduced by Bernstein (1927). This consists of eliminating from the original dependent sequence several blocks of contiguous points, leaving us with some remaining blocks of points. Instead of these dependent blocks, we then consider independent blocks of points, each with the same size and the same distribution (within each block) as the dependent ones. By Lemma 3, for a β-mixing distribution, the expected value of a random variable defined over the dependent blocks is close to the one based on these independent blocks. Working with these independent blocks brings us back to a situation similar to the i.i.d. case, with i.i.d. blocks replacing i.i.d. points. Figure 1 illustrates the two types of blocks just discussed.

Our use of this method somewhat differs from previous ones (see Yu, 1994; Meir, 2000) where many blocks of equal size are considered. We will be dealing with four blocks and with typically unequal sizes. More specifically, note that for Equation 2 to hold, we only need that the variable Z_i be independent of the other points in the sequence. To achieve this, roughly speaking, we will be "discarding" some of the points in the sequence surrounding Z_i. This results in a sequence of three blocks of contiguous points. If our algorithm is stable and we do not discard too many points, the hypothesis returned should not be greatly affected by this operation. In the next step, we apply the independent block lemma, which then allows us to assume each of these blocks as independent modulo the addition of a mixing term. In particular, Z_i becomes independent of all other points. Clearly, the number of points discarded is subject to a trade-off: removing too many points could excessively modify the hypothesis returned; removing too few would maintain the dependency between Z_i and the remaining points, thereby inducing a larger penalty when applying Lemma 3. This trade-off is made explicit in the following section where an optimal solution is sought.

3.1 Lipschitz Bound

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 $\widetilde{R}(h_{S_b}) = \mathrm{E}_{\widetilde{z}}[c(h_{S_b}, \widetilde{z})]$ the expectation where 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 2(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 5 *Assume that the learning algorithm is $\widehat{\beta}$-stable and that the cost function c is bounded by M. Then, for any sample S of size m drawn from a φ-mixing stationary distribution and for any $b \in \{0, \ldots, m\}$, the following holds:*

$$|R(h_S) - \widetilde{R}(h_{S_b})| \leq b\widehat{\beta} + M\varphi(b)$$

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

$$|R(h_S) - R(h_{S_b})| = |\operatorname*{E}_z[c(h_S, z)|S] - \operatorname*{E}_z[c(h_{S_b}, z)|S_b]| \leq b\widehat{\beta}. \tag{3}$$

Now, in order to remove the dependence on S_b we bound the following difference

$$\begin{aligned}
&|\operatorname*{E}_z[c(h_{S_b}, z)|S_b] - \operatorname*{E}_{\widetilde{z}}[c(h_{S_b}, \widetilde{z})]| \\
&\qquad = \Big|\sum_{z \in Z} c(h_{S_b}, z)(\Pr[z|S_b] - \Pr[z])\Big| \\
&\qquad = \Big|\sum_{z \in Z^+} c(h_{S_b}, z)(\Pr[z|S_b] - \Pr[z]) + \sum_{z \in Z^-} c(h_{S_b}, z)(\Pr[z|S_b] - \Pr[z])\Big| \\
&\qquad = \Big|\sum_{z \in Z^+} c(h_{S_b}, z)\Big|\Pr[z|S_b] - \Pr[z]\Big| - \sum_{z \in Z^-} c(h_{S_b}, z)\Big|\Pr[z|S_b] - \Pr[z]\Big|\Big| \\
&\qquad \leq \max_{Z \in \{Z^-, Z^+\}} \sum_{z \in Z} c(h_{S_b}, z)\Big|\Pr[z|S_b] - \Pr[z]\Big| \qquad (4) \\
&\qquad \leq \max_{Z \in \{Z^-, Z^+\}} M \sum_{z \in Z} \Big|\Pr[z|S_b] - \Pr[z]\Big| \\
&\qquad = \max_{Z \in \{Z^-, Z^+\}} M \Big|\sum_{z \in Z} \Pr[z|S_b] - \Pr[z]\Big| \\
&\qquad = \max_{Z \in \{Z^-, Z^+\}} M \Big|\Pr[Z|S_b] - \Pr[Z]\Big| \leq M\varphi(b),
\end{aligned}$$

where the sum has been separated over the set of zs Z^+ for which the difference $\Pr[z|S_b] - \Pr[z]$ is non-negative, and its complement Z^-. Using (3) and (4) and the triangle inequality yields the statement of the lemma. ■

Note that we assume that z immediately follows the sample S, which is the strongest dependent scenario. The following bounds can be improved in a straightforward manner if the test point z is assumed to be observed say k units of time after the sample S. The bound would then contain the mixing term $\varphi(k + b)$ instead of $\varphi(b)$.

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

Lemma 6 *Let $S = (z_1, \ldots, z_i, \ldots, z_m)$ and $S^i = (z_1, \ldots, z_i', \ldots, z_m)$ be two sequences drawn from a φ-mixing stationary process that differ only in point z_i for some $i \in \{1, \ldots, m\}$, and let h_S and h_{S^i} be the hypotheses returned by a $\widehat{\beta}$-stable algorithm when trained on each of these samples. Then, for any $i \in \{1, \ldots, m\}$, the following inequality holds:*

$$|\Phi(S) - \Phi(S^i)| \leq (b + 2)2\widehat{\beta} + 2\varphi(b)M + \frac{M}{m}.$$

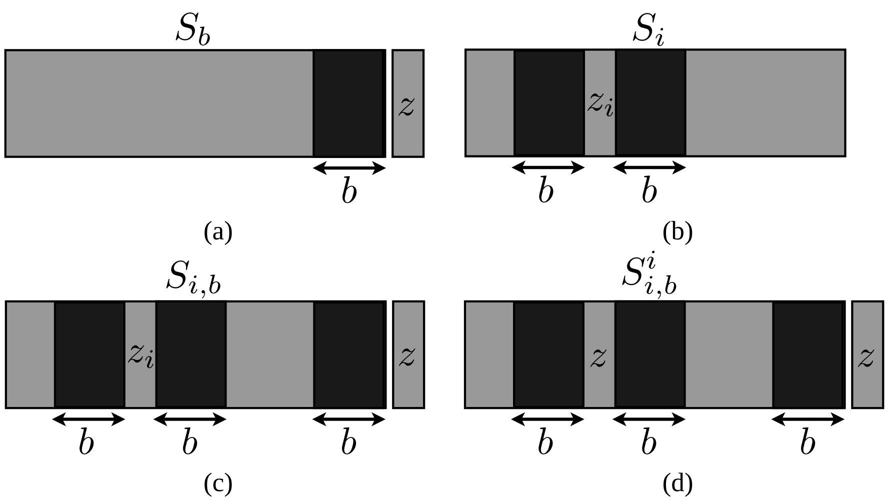

Figure 2: Illustration of the sequences derived from S that are considered in the proofs.

Proof To prove this inequality, we first bound the difference of the empirical errors as in Bousquet and Elisseeff (2002), then the difference of the generalization errors. Bounding the difference of costs on agreeing points with $\widehat{\beta}$ and the one that disagrees with M gives

$$\begin{aligned}|\widehat{R}(h_S)-\widehat{R}(h_{S^i})| &\leq \frac{1}{m}\sum_{j\neq i}|c(h_S,z_j)-c(h_{S^i},z_j)|+\frac{1}{m}|c(h_S,z_i)-c(h_{S^i},z_i')|\\ &\leq 2\widehat{\beta}+\frac{M}{m}. \end{aligned} \tag{5}$$

Since both $R(h_S)$ and $R(h_{S^i})$ are defined with respect to a (different) dependent point, we can apply Lemma 5 to both generalization error terms and use $\widehat{\beta}$-stability. Using this and the triangle inequality, we can write

$$\begin{aligned}|R(h_S)-R(h_{S^i})| &\leq |R(h_S)-\widetilde{R}(h_{S_b})+\widetilde{R}(h_{S_b})-\widetilde{R}(h_{S_b^i})+\widetilde{R}(h_{S_b^i})-R(h_{S^i})|\\ &\leq |\widetilde{R}(h_{S_b})-\widetilde{R}(h_{S_b^i})|+2b\widehat{\beta}+2\varphi(b)M\\ &= \mathop{\mathrm{E}}_{\widetilde{z}}[c(h_{S_b},\widetilde{z})-c(h_{S_b^i},\widetilde{z})]+2b\widehat{\beta}+2\varphi(b)M\\ &\leq 2\widehat{\beta}+2b\widehat{\beta}+2\varphi(b)M. \end{aligned} \tag{6}$$

The statement of the lemma is obtained by combining inequalities 5 and 6. ■

3.2 Bound on Expectation

As mentioned earlier, to obtain an explicit bound after application of a generalized McDiarmid's inequality, we also need to bound $\mathrm{E}_S[\Phi(S)]$. This is done by analyzing independent blocks using Lemma 3.

Lemma 7 *Let h_S be the hypothesis returned by a $\widehat{\beta}$-stable algorithm trained on a sample S drawn from a β-mixing stationary distribution. Then, for all $b \in [1, m]$, the following inequality holds:*

$$\mathop{\mathrm{E}}_{S}[|\Phi(S)|] \leq (6b+2)\widehat{\beta} + 3\beta(b)M.$$

Proof Let S_b be defined as in the proof of Lemma 5. 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 2(a-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 as seen in Figure 2(d).

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

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

The application of Lemma 3 to the difference of two cost functions also bounded by M as in the right-hand side leads to

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

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

$$\mathop{\mathrm{E}}_{S}[\Phi(S)] \leq \mathop{\mathrm{E}}_{\widetilde{S}_{i,b},\widetilde{z}}\left[\frac{1}{m}\sum_{i=1}^{m} c(h_{\widetilde{S}^i_{i,b}}, \widetilde{z}) - c(h_{\widetilde{S}_{i,b}}, \widetilde{z})\right] + 6b\widehat{\beta} + 3\beta(b)M \leq 2\widehat{\beta} + 6b\widehat{\beta} + 3\beta(b)M,$$

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 $\widehat{\beta}$-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 φ-Mixing Concentration Bound

We are now prepared to make use of a concentration inequality to provide a generalization bound in the φ-mixing scenario. Several concentration inequalities have been shown in the φ-mixing case, for example, Marton (1998), Samson (2000), Chazottes et al. (2007) and Kontorovich and Ramanan (2008). We will use that of Kontorovich and Ramanan (2008), which is very similar to that of Chazottes et al. (2007), modulo the fact that the latter requires a finite sample space.

The following is a concentration inequality derived from that of Kontorovich and Ramanan (2008).[3]

Theorem 8 *Let $\Phi\colon Z^m \to \mathbb{R}$ be a measurable function that is c-Lipschitz with respect to the Hamming metric for some $c > 0$ and let $Z_1, \ldots, Z_m$ be random variables distributed according to a φ-mixing distribution. Then, for any $\epsilon > 0$, the following inequality holds:*

$$\Pr\Big[\big|\Phi(Z_1,\ldots,Z_m) - \mathrm{E}[\Phi(Z_1,\ldots,Z_m)]\big| \geq \epsilon\Big] \leq 2\exp\left(\frac{-2\epsilon^2}{mc^2||\Delta_m||_\infty^2}\right),$$

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

It should be pointed out that the statement of the theorem in this paper is improved by a factor of 4 in the exponent with respect to that of Kontorovich and Ramanan (2008, Theorem 1.1). This can be achieved straightforwardly by following the same steps as in the proof of Kontorovich and Ramanan (2008), but by making use of the following general form of McDiarmid's inequality (Theorem 9) instead of Azuma's inequality. In particular, Theorem 5.1 of Kontorovich and Ramanan (2008) shows that for a φ-mixing distribution and a 1-Lipschitz function, the constants c_i can be bounded as follows in Theorem 9:

$$c_i \leq 1 + 2\sum_{k=1}^{m-i}\varphi(k).$$

Theorem 9 (McDiarmid, 1989, 6.10) *Let $Z_1, \ldots, Z_m$ be arbitrary random variables taking values in Z and let $\Phi\colon Z^m \to \mathbb{R}$ be a measurable function satisfying for all $z_i, z_i' \in Z$, $i = 1, \ldots, m$, the following inequalities:*

$$\left|\mathrm{E}\Big[\Phi(Z_1,\ldots,Z_m)\big|Z_1 = z_1,\ldots,Z_i = z_i\Big] - \mathrm{E}\Big[\Phi(Z_1,\ldots,Z_m)\big|Z_1 = z_1,\ldots,Z_i = z_i'\Big]\right| \leq c_i,$$

where $c_i > 0$, $i = 1, \ldots, m$, are constants. Then, for any $\epsilon > 0$, the following inequality holds :

$$\Pr\Big[\big|\Phi(Z_1,\ldots,Z_m) - \mathrm{E}[\Phi(Z_1,\ldots,Z_m)]\big| \geq \epsilon\Big] \leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right).$$

In the i.i.d. case, McDiarmid's theorem can be restated in the following simpler form that we shall use in Section 4.

Theorem 10 (McDiarmid, i.i.d. scenario) *Let $Z_1, \ldots, Z_m$ be independent random variables taking values in Z and let $\Phi\colon Z^m \to \mathbb{R}$ be a measurable function satisfying for all $z_i, z_i' \in Z$, $i = 1, \ldots, m$, the following inequalities:*

$$\big|\Phi(z_1,\ldots,z_i,\ldots z_m) - \Phi(z_1,\ldots,z_i',\ldots z_m)\big| \leq c_i,$$

where $c_i > 0$, $i = 1, \ldots, m$, are constants. Then, for any $\epsilon > 0$, the following inequality holds:

$$\Pr\Big[\big|\Phi(Z_1,\ldots,Z_m) - \mathrm{E}[\Phi(Z_1,\ldots,Z_m)]\big| \geq \epsilon\Big] \leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right).$$

3. We should note that original bound is expressed in terms of η-mixing coefficients. To simplify presentation, we are adapting it to the case of stationary φ-mixing sequences by using the following straightforward inequality for a stationary process: $2\varphi(j-i) \geq \eta_{ij}$. Furthermore, the bound presented in Kontorovich and Ramanan (2008) holds when the sample space is countable, it is extended to the continuous case in Kontorovich (2007).

3.4 φ-Mixing Generalization Bounds

This section presents several theorems that constitute the main results of this paper in the φ-mixing case. The following theorem is constructed from the bounds shown in the previous three sections.

Theorem 11 (General Non-i.i.d. Stability Bound) *Let h_S denote the hypothesis returned by a $\widehat{\beta}$-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 $b \in \{0, \ldots, m\}$ and any $\varepsilon > 0$, the following generalization bound holds:*

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

Proof The theorem follows directly the application of Lemma 6 and Lemma 7 to Theorem 8. ■

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 12 (Non-i.i.d. Stability Bound for Algebraically Mixing Sequences) *Let h_S denote the hypothesis returned by a $\widehat{\beta}$-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 $\varepsilon > 0$, the following generalization bound holds:*

$$\Pr_S\left[\left|R(h_S) - \widehat{R}(h_S)\right| > \varepsilon + 8\widehat{\beta} + (r+1)6M\varphi(b)\right] \leq 2\exp\left(\frac{-2\varepsilon^2(1+2\varphi_0 r/(r-1))^{-2}}{m(6\widehat{\beta} + (r+1)2M\varphi(b) + M/m)^2}\right),$$

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

Proof For an algebraically mixing sequence, the value of b minimizing the bound of Theorem 11 satisfies the equation $\widehat{\beta} b^* = rM\varphi(b^*)$. Since b must be an integer, we use the approximation $b = \left\lceil \left(\frac{\widehat{\beta}}{r\varphi_0 M}\right)^{-1/(r+1)} \right\rceil$ when applying Theorem 11. However, observing the inequalities $\varphi(b^*) \geq \varphi(b)$ and $(b^*+1) \geq b$, allows us to write the statement of Theorem 12 in terms of the fractional choice b^*.

The term in the numerator can be bounded as

$$\begin{aligned}
1 + 2\sum_{i=1}^m \varphi(i) &= 1 + 2\sum_{i=1}^m \varphi_0 i^{-r} \\
&\leq 1 + 2\varphi_0\left(1 + \int_1^m x^{-r}dx\right) \\
&= 1 + 2\varphi_0\left(1 + \frac{m^{1-r}-1}{1-r}\right).
\end{aligned}$$

Using the assumption $r > 1$, we can upper bound m^{1-r} with 1 and obtain

$$1+2\varphi_0\left(1+\frac{m^{1-r}-1}{1-r}\right) \leq 1+2\varphi_0\left(1+\frac{1}{r-1}\right) = 1+\frac{2\varphi_0 r}{r-1}.$$

Plugging in this value and the minimizing value of b in the bound of Theorem 11 yields the statement of the theorem. ■

In the case of a zero mixing coefficient ($\varphi = 0$ and $b = 0$), the bounds of Theorem 11 coincide with the i.i.d. stability bound of Bousquet and Elisseeff (2002).

In the general case, in order for the right-hand side of these bounds to converge, we must have $\widehat{\beta} = o(1/\sqrt{m})$ and $\varphi(b) = o(1/\sqrt{m})$. The first condition holds for several families of algorithms with $\widehat{\beta} \leq O(1/m)$ (Bousquet and Elisseeff, 2002).

In the case of algebraically mixing sequences with $r > 1$, as assumed in Theorem 12, $\widehat{\beta} \leq O(1/m)$ implies $\varphi(b) \approx \varphi_0(\widehat{\beta}/(r\varphi_0 M))^{(r/(r+1))} < O(1/\sqrt{m})$. More specifically, for the scenario of algebraic mixing with $1/m$-stability, the following bound holds with probability at least $1-\delta$:

$$\left|R(h_S) - \widehat{R}(h_S)\right| \leq O\left(\sqrt{\frac{\log(1/\delta)}{m^{\frac{r-1}{r+1}}}}\right).$$

This is obtained by setting the right-hand side of Theorem 12 equal to δ and solving for ε. Furthermore, if we choose $\varepsilon = \sqrt{\frac{C\log(m)}{m^{(r-1)/(r+1)}}}$ for a large enough constant $C > 0$, the right-hand side of Theorem 12 is summable over m and thus, by the Borel-Cantelli lemma, the following inequality holds almost surely:

$$\left|R(h_S) - \widehat{R}(h_S)\right| \leq O\left(\sqrt{\frac{\log(m)}{m^{\frac{r-1}{r+1}}}}\right).$$

Similar bounds can be given for the exponential mixing setting ($\varphi(k) = \varphi_0 \exp(-\varphi_1 k^r)$). If we choose $b = O(\sqrt{\log(m)^3/m})$ and assume $\widehat{\beta} = O(1/m)$, then, with probability at least $1-\delta$,

$$\left|R(h_S) - \widehat{R}(h_S)\right| \leq O\left(\sqrt{\frac{\log(1/\delta)\log^2(m)}{m}}\right).$$

If we instead set $\varepsilon = C\sqrt{\frac{\log^3(m)}{m}}$ for a large enough constant C, the right-hand side of Theorem 12 is summable and again by the Borel-Cantelli lemma we have

$$\left|R(h_S) - \widehat{R}(h_S)\right| \leq O\left(\sqrt{\frac{\log^3(m)}{m}}\right),$$

almost surely.

3.5 Applications

We now present the application of our stability bounds for algebraically φ-mixing sequences to several algorithms, including the family of kernel-based regularization algorithms and that of relative entropy-based regularization algorithms. The application of our learning bounds will benefit from the previous analysis of the stability of these algorithms by Bousquet and Elisseeff (2002).

3.5.1 KERNEL-BASED REGULARIZATION ALGORITHMS

We first apply our bounds to a family of algorithms minimizing a regularized objective function based on the norm $\|\cdot\|_K$ in a reproducing kernel Hilbert space, where K is a positive definite symmetric kernel:

$$\underset{h\in H}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m} c(h,z_i) + \lambda\|h\|_K^2. \tag{7}$$

The application of our bound is possible, under some general conditions, since kernel regularized algorithms are stable with $\widehat{\beta} \le O(1/m)$ (Bousquet and Elisseeff, 2002). For the sake of completeness, we briefly present the proof of this $\widehat{\beta}$-stability.

We will assume that the cost function c is σ-*admissible*, that is there exists $\sigma \in \mathbb{R}_+$ such that for any two hypotheses $h,h' \in H$ and for all $z=(x,y)\in X\times Y$,

$$|c(h,z)-c(h',z)| \le \sigma|h(x)-h'(x)|.$$

This assumption holds for the quadratic cost and most other cost functions when the hypothesis set and the set of output labels are bounded by some $M \in \mathbb{R}_+$: $\forall h \in H, \forall x \in X, |h(x)| \le M$ and $\forall y \in Y, |y| \le M$. We will also assume that c is differentiable. This assumption is in fact not necessary and all of our results hold without it, but it makes the presentation simpler.

We denote by B_F the Bregman divergence associated to a convex function F: $B_F(f\|g) = F(f) - F(g) - \langle f-g, \nabla F(g)\rangle$. In what follows, it will be helpful to define F as the objective function of a general regularization based algorithm,

$$F_S(h) = \widehat{R}_S(h) + \lambda N(h),$$

where $\widehat{R}_S$ is the empirical error as measured on the sample S, $N: H \to \mathbb{R}^+$ is a regularization function and $\lambda > 0$ is the familiar trade-off parameter. Finally, we shall use the shorthand $\Delta h = h' - h$.

Lemma 13 (Bousquet and Elisseeff, 2002) *A kernel-based regularization algorithm of the form (7), with bounded kernel $K(x,x) \le \kappa < \infty$ and σ-admissible cost function, is $\widehat{\beta}$-stable with coefficient*

$$\widehat{\beta} \le \frac{\sigma^2\kappa^2}{m\lambda}.$$

Proof Let h and h' be the minimizers of F_S and F'_S respectively where S and S' differ in the first coordinate (choice of coordinate is without loss of generality), then,

$$B_N(h'\|h) + B_N(h\|h') \le \frac{2\sigma}{m\lambda}\sup_{x\in S}|\Delta h(x)|. \tag{8}$$

To see this, we notice that since $B_F = B_{\widehat{R}} + \lambda B_N$, and since a Bregman divergence is non-negative,

$$\lambda\big(B_N(h'\|h) + B_N(h\|h')\big) \le B_{F_S}(h'\|h) + B_{F_{S'}}(h\|h').$$

By the definition of h and h' as the minimizers of F_S and $F_{S'}$,

$$B_{F_S}(h'\|h) + B_{F_{S'}}(h\|h') = \widehat{R}_{F_S}(h') - \widehat{R}_{F_S}(h) + \widehat{R}_{F_{S'}}(h) - \widehat{R}_{F_{S'}}(h').$$

Finally, by the σ-admissibility of the cost function c and the definition of S and S',

$$\begin{aligned}\lambda\big(B_N(h'\|h)+B_N(h\|h')\big) &\leq \widehat{R}_{F_S}(h')-\widehat{R}_{F_S}(h)+\widehat{R}_{F_{S'}}(h)-\widehat{R}_{F_{S'}}(h') \\ &= \frac{1}{m}\Big[c(h',z_1)-c(h,z_1)+c(h,z_1')-c(h',z_1')\Big] \\ &\leq \frac{1}{m}\Big[\sigma|\Delta h(x_1)|+\sigma|\Delta h(x_1')|\Big] \\ &\leq \frac{2\sigma}{m}\sup_{x\in S}|\Delta h(x)|,\end{aligned}$$

which establishes (8).

Now, if we consider $N(\cdot)=\|\cdot\|_K^2$, we have $B_N(h'\|h)=\|h'-h\|_K^2$, thus $B_N(h'\|h)+B_N(h\|h')=2\|\Delta h\|_K^2$ and by (8) and the reproducing kernel property,

$$\begin{aligned}2\|\Delta h\|_K^2 &\leq \frac{2\sigma}{m\lambda}\sup_{x\in S}|\Delta h(x)| \\ &\leq \frac{2\sigma}{m\lambda}\kappa||\Delta h||_K.\end{aligned}$$

Thus $\|\Delta h\|_K \leq \frac{\sigma\kappa}{m\lambda}$. And using the σ-admissibility of c and the kernel reproducing property we obtain

$$\forall z\in X\times Y, |c(h',z)-c(h,z)|\leq \sigma|\Delta h(x)|\leq \kappa\sigma\|\Delta h\|_K.$$

Therefore,

$$\forall z\in X\times Y, |c(h',z)-c(h,z)|\leq \frac{\sigma^2\kappa^2}{m\lambda},$$

which completes the proof. ■

Three specific instances of kernel regularization algorithms are SVR, for which the cost function is based on the ε-insensitive cost:

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

Kernel Ridge Regression (Saunders et al., 1998), for which

$$c(h,z)=(h(x)-y)^2\ ,$$

and finally Support Vector Machines with the hinge-loss,

$$c(h,z)=\begin{cases}0 & \text{if } 1-yh(x)\leq 0, \\ 1-yh(x) & \text{if } yh(x)<1,\end{cases}$$

For kernel regularization algorithms, as pointed out in Bousquet and Elisseeff (2002, Lemma 23), a bound on the labels immediately implies a bound on the output of the hypothesis returned by the algorithm. We formally state this lemma below.

Lemma 14 *Let h^* be the solution of the optimization problem (7), let c be a cost function and let $B(\cdot)$ be a real-valued function such that for all $h \in H$, $x \in X$, and $y' \in Y$,*

$$c(h(x), y') \leq B(h(x)).$$

Then, the output of h^* is bounded as follows,

$$\forall x \in X, |h^*(x)| \leq \kappa \sqrt{\frac{B(0)}{\lambda}},$$

where λ is the regularization parameter, and $\kappa^2 \geq K(x,x)$ for all $x \in X$.

Proof Let $F(h) = \frac{1}{m}\sum_{i=1}^m c(h, z_i) + \lambda \|h\|_K^2$ and let $\mathbf{0}$ be the zero hypothesis, then by definition of F and h^*,

$$\lambda \|h^*\|_K^2 \leq F(h^*) \leq F(\mathbf{0}) \leq B(0).$$

Then, using the reproducing kernel property and the Cauchy-Schwarz inequality we note,

$$\forall x \in X, |h^*(x)| = \langle h^*, K(x, \cdot)\rangle \leq \|h^*\|_K \sqrt{K(x,x)} \leq \kappa \|h^*\|_K.$$

Combining the two inequalities proves the lemma. ■

We note that in Bousquet and Elisseeff (2002), the following bound is also stated: $c(h^*(x), y') \leq B(\kappa\sqrt{B(0)/\lambda})$. However, when later applied, it seems that the authors use an incorrect upper bound function $B(\cdot)$, which we remedy in the following.

Corollary 15 *Assume a bounded output $Y = [0, B]$, for some $B > 0$, and assume that $K(x,x) \leq \kappa^2$ 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. Let $u = r/(r+1) \in [\frac{1}{2}, 1]$, $M' = 2(r+1)\varphi_0 M/(r\varphi_0 M)^u$, and $\varphi_0' = (1 + 2\varphi_0 r/(r-1))$. Then, with probability at least $1-\delta$, the following generalization bounds hold for*

a. *Support Vector Machines (SVM, with hinge-loss)*

$$R(h_S) \leq \widehat{R}(h_S) + \frac{8\kappa^2}{\lambda m} + \left(\frac{2\kappa^2}{\lambda}\right)^u \frac{3M'}{m^u} + \varphi_0'\left(M + \frac{3\kappa^2}{\lambda} + \left(\frac{2\kappa^2}{\lambda}\right)^u \frac{M'}{m^{u-1}}\right)\sqrt{\frac{2\log(2/\delta)}{m}},$$

where $M = \kappa\sqrt{\frac{1}{\lambda}} + B$.

b. *Support Vector Regression (SVR):*

$$R(h_S) \leq \widehat{R}(h_S) + \frac{8\kappa^2}{\lambda m} + \left(\frac{2\kappa^2}{\lambda}\right)^u \frac{3M'}{m^u} + \varphi_0'\left(M + \frac{3\kappa^2}{\lambda} + \left(\frac{2\kappa^2}{\lambda}\right)^u \frac{M'}{m^{u-1}}\right)\sqrt{\frac{2\log(2/\delta)}{m}},$$

where $M = \kappa\sqrt{\frac{2B}{\lambda}} + B$.

c. Kernel Ridge Regression (KRR):

$$R(h_S) \leq \widehat{R}(h_S) + \frac{32\kappa^2 B^2}{\lambda m} + \left(\frac{8\kappa^2 B^2}{\lambda}\right)^u \frac{3M'}{m^u} + \varphi_0'\left(M + \frac{12\kappa^2 B^2}{\lambda} + \left(\frac{8\kappa^2 B^2}{\lambda}\right)^u \frac{M'}{m^{u-1}}\right)\sqrt{\frac{2\log(2/\delta)}{m}},$$

where $M = 2\kappa^2 B^2/\lambda + B^2$.

Proof For SVM, the hinge-loss is 1-admissible giving $\widehat{\beta} \leq \kappa^2/(\lambda m)$. Using Lemma 14, with $B(0) = 1$, the loss can be bounded $\forall x \in X, y \in Y, 1 + |h^*(x)| \leq \kappa\sqrt{\frac{1}{\lambda}} + B$.

Similarly, SVR has a loss function that is 1-admissible, thus, applying Lemma 13 gives us $\widehat{\beta} \leq \kappa^2/(\lambda m)$. Using Lemma 14, with $B(0) = B$, we can bound the loss as follows, $\forall x \in X, y \in Y, |h^*(x) - y| \leq \kappa\sqrt{\frac{B}{\lambda}} + B$.

Finally for KRR, we have a loss function that is $2B$-admissible and again using Lemma 13 $\widehat{\beta} \leq 4\kappa^2 B^2/(\lambda m)$. Again, applying Lemma 14 with $B(0) = B^2$ and $\forall x \in X, y \in Y, (h^*(x) - y)^2 \leq \kappa^2 B^2/\lambda + B^2$.

Plugging these values into the bound of Theorem 12 and setting the right-hand side to δ yields the statement of the corollary. ■

3.5.2 Relative Entropy-Based Regularization Algorithms

In this section, we apply the results of Theorem 12 to a family of learning algorithms based on relative entropy-regularization. These algorithms learn hypotheses h that are mixtures of base hypotheses in $\{h_\theta : \theta \in \Theta\}$, where Θ is measurable set. The output of these algorithms is a mixture $g : \Theta \to \mathbb{R}$, that is a distribution over Θ. Let G denote the set of all such distributions and let $g_0 \in G$ be a fixed distribution. Relative entropy based-regularization algorithms output the solution of a minimization problem of the following form:

$$\operatorname*{argmin}_{g \in G} \frac{1}{m} \sum_{i=1}^{m} c(g, z_i) + \lambda D(g \| g_0), \tag{9}$$

where the cost function $c : G \times Z \to \mathbb{R}$ is defined in terms of a second internal cost function $c' : H \times Z \to \mathbb{R}$:

$$c(g, z) = \int_\Theta c'(h_\theta, z) g(\theta) d\theta,$$

and where $D(g \| g_0)$ is the relative entropy between g and g_0:

$$D(g \| g_0) = \int_\Theta g(\theta) \log \frac{g(\theta)}{g_0(\theta)} d\theta.$$

As shown by Bousquet and Elisseeff (2002, Theorem 24), a relative entropy-based regularization algorithm defined by (9) with bounded loss $c'(\cdot) \leq M$, is $\widehat{\beta}$-stable with the following bound on the stability coefficient:

$$\widehat{\beta} \leq \frac{M^2}{\lambda m}.$$

Theorem 12 combined with this inequality immediately yields the following generalization bound.

Corollary 16 *Let h_S be the hypothesis solution of the optimization (9) trained on a sample S drawn from an algebraically φ-mixing stationary distribution with the internal cost function c' bounded by M. Then, with probability at least $1-\delta$, the following holds:*

$$R(h_S) \leq \widehat{R}(h_S) + \frac{8M^2}{\lambda m} + \frac{3M'}{\lambda^u m^u} + \varphi_0' \left(M + \frac{3M^2}{\lambda} + \frac{2^u M'}{\lambda^u m^{u-1}} \right) \sqrt{\frac{2\log(2/\delta)}{m}},$$

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

3.6 Discussion

The results presented here are, to the best of our knowledge, the first stability-based generalization bounds for the class of algorithms just studied in a non-i.i.d. scenario. These bounds are non-trivial when the condition on the regularization parameter $\lambda \gg 1/m^{1/2-1/r}$ parameter holds for all large values of m. This condition coincides with the one obtained in the i.i.d. setting by Bousquet and Elisseeff (2002), in the limit, as r tends to infinity. The next section gives stability-based generalization bounds that hold even in the scenario of β-mixing sequences.

4. β-Mixing Generalization Bounds

In this section, we prove a stability-based generalization bound that only requires the training sequence to be drawn from a β-mixing stationary distribution. The bound is thus more general and covers the φ-mixing case analyzed in the previous section. However, unlike the φ-mixing case, the β-mixing bound presented here is not a purely exponential bound. It contains an additive term, which depends on the mixing coefficient.

As in the previous section, $\Phi(S)$ is defined by $\Phi(S) = R(h_S) - \widehat{R}(h_S)$. To simplify the presentation, here, we define the generalization error of h_S by $R(h_S) = \mathrm{E}_z[c(h_S, z)]$. Thus, test samples are assumed independent of S.[4] Note that for any block of points $Z = z_1 \ldots z_k$ drawn independently of S, the following equality holds:

$$\mathop{\mathrm{E}}_{Z}\left[\frac{1}{|Z|}\sum_{z\in Z} c(h_S, z)\right] = \frac{1}{k}\sum_{i=1}^{k} \mathop{\mathrm{E}}_{Z}[c(h_S, z_i)] = \frac{1}{k}\sum_{i=1}^{k} \mathop{\mathrm{E}}_{z_i}[c(h_S, z_i)] = \mathop{\mathrm{E}}_{z}[c(h_S, z)]$$

since, by stationarity, $\mathrm{E}_{z_i}[c(h_S, z_i)] = \mathrm{E}_{z_j}[c(h_S, z_j)]$ for all $1 \leq i, j \leq k$. Thus, for any such block Z, we can write $R(h_S) = \mathrm{E}_Z\left[\frac{1}{|Z|}\sum_{z\in Z} c(h_S, z)\right]$. For convenience, we extend the cost function c to blocks as follows:

$$c(h, Z) = \frac{1}{|Z|}\sum_{z\in Z} c(h, z).$$

With this notation, $R(h_S) = \mathrm{E}_Z[c(h_S, Z)]$ for any block drawn independently of S, regardless of the size of Z.

To derive a generalization bound for the β-mixing scenario, we apply McDiarmid's inequality (Theorem 10) to Φ defined over a sequence of independent blocks. The independent blocks we consider are non-symmetric and thus more general than those considered by previous authors (Yu, 1994; Meir, 2000).

4. In the β-mixing scenario, a result similar to that of Lemma 5 can be shown to hold in expectation with respect the sample S. Using Markov's inequality, the inequality can be shown to hold with high probability. Thus, the results that follow can all be be extended to the case where the test points depend on the training sample, at the expense of an an additional confidence term.

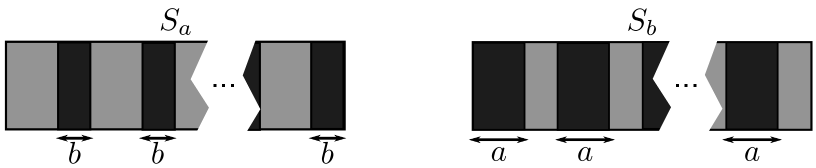

Figure 3: Illustration of the sequences S_a and S_b derived from S that are considered in the proofs. The darkened regions are considered as being removed from the sequence.

From a sample S made of a sequence of m points, we construct two sequences of blocks S_a and S_b, each containing μ blocks. Each block in S_a contains a points and each block in S_b contains b points (see Figure 3). S_a and S_b form a partitioning of S; for any $a, b \in \{0, \ldots, m\}$ such that $(a+b)\mu = m$, they are defined precisely as follows:

$$S_a = (Z_1^{(a)}, \ldots, Z_\mu^{(a)}), \text{ with } Z_i^{(a)} = z_{(i-1)(a+b)+1}, \ldots, z_{(i-1)(a+b)+a}$$
$$S_b = (Z_1^{(b)}, \ldots, Z_\mu^{(b)}), \text{ with } Z_i^{(b)} = z_{(i-1)(a+b)+a+1}, \ldots, z_{(i-1)(a+b)+a+b},$$

for all $i \in \{1, \ldots, \mu\}$. We shall consider similarly sequences of i.i.d. blocks $\widetilde{Z}_i^a$ and $\widetilde{Z}_i^b$, $i \in \{1, \ldots, \mu\}$, such that the points within each block are drawn according to the same original β-mixing distribution and shall denote by $\widetilde{S}_a$ the block sequence $(\widetilde{Z}_1^{(a)}, \ldots, \widetilde{Z}_\mu^{(a)})$.

In preparation for the application of McDiarmid's inequality, we give a bound on the expectation of $\Phi(\widetilde{S}_a)$. Since the expectation is taken over a sequence of i.i.d. blocks, this brings us to a situation similar to the i.i.d. scenario analyzed by Bousquet and Elisseeff (2002), with the exception that we are dealing with i.i.d. blocks instead of i.i.d. points.

Lemma 17 *Let $\widetilde{S}_a$ be an independent block sequence as defined above, then the following bound holds for the expectation of $|\Phi(\widetilde{S}_a)|$:*

$$\mathop{\mathrm{E}}_{\widetilde{S}_a}[|\Phi(\widetilde{S}_a)|] \leq 2a\widehat{\beta}.$$

Proof Since the blocks $\widetilde{Z}^{(a)}$ are independent, we can replace any one of them with any other block Z drawn from the same distribution. However, changing the training set also changes the hypothesis, in a limited way. This is shown precisely below:

$$\begin{aligned}
\mathop{\mathrm{E}}_{\widetilde{S}_a}[|\Phi(\widetilde{S}_a)|] &= \mathop{\mathrm{E}}_{\widetilde{S}_a}\left[\left|\frac{1}{\mu}\sum_{i=1}^{\mu} c(h_{\widetilde{S}_a}, \widetilde{Z}_i^{(a)}) - \mathop{\mathrm{E}}_{Z}[c(h_{\widetilde{S}_a}, Z)]\right|\right] \\
&\leq \mathop{\mathrm{E}}_{\widetilde{S}_a, Z}\left[\left|\frac{1}{\mu}\sum_{i=1}^{\mu} c(h_{\widetilde{S}_a}, \widetilde{Z}_i^{(a)}) - c(h_{\widetilde{S}_a}, Z)\right|\right] \\
&= \mathop{\mathrm{E}}_{\widetilde{S}_a, Z}\left[\left|\frac{1}{\mu}\sum_{i=1}^{\mu} c(h_{\widetilde{S}_a^i}, Z) - c(h_{\widetilde{S}_a}, Z)\right|\right],
\end{aligned}$$

where $\widetilde{S}_a^i$ corresponds to the block sequence $\widetilde{S}_a$ obtained by replacing the ith block with Z. The $\widehat{\beta}$-stability of the learning algorithm gives

$$\mathop{\mathrm{E}}_{\widetilde{S}_a,Z}\left[\frac{1}{\mu}\Big|\sum_{i=1}^{\mu} c(h_{\widetilde{S}_a^i},Z)-c(h_{\widetilde{S}_a},Z)\Big|\right] \leq \mathop{\mathrm{E}}_{\widetilde{S}_a,Z}\left[\frac{1}{\mu}\sum_{i=1}^{\mu} 2a\widehat{\beta}\right] \leq 2a\widehat{\beta}.$$

■

We now relate the non-i.i.d. event $\Pr[\Phi(S)\geq\epsilon]$ to an independent block sequence event to which we can apply McDiarmid's inequality.

Lemma 18 *Assume a $\widehat{\beta}$-stable algorithm. Then, for a sample S drawn from a β-mixing stationary distribution, the following bound holds:*

$$\Pr_S[|\Phi(S)|\geq\epsilon] \leq \Pr_{\widetilde{S}_a}\left[|\Phi(\widetilde{S}_a)|-\mathrm{E}[|\Phi(\widetilde{S}_a)|]\geq\epsilon_0'\right]+(\mu-1)\beta(b),$$

where $\epsilon_0' = \epsilon - \frac{\mu bM}{m} - 2\mu b\widehat{\beta} - \mathrm{E}_{\widetilde{S}_a'}[|\Phi(\widetilde{S}_a')|]$.

Proof The proof consists of first rewriting the event in terms of S_a and S_b and bounding the error on the points in S_b in a trivial manner. This can be afforded since b will be eventually chosen to be small. Since $|\mathrm{E}_{Z'}[c(h_S,Z')]-c(h_S,z')|\leq M$ for any $z'\in S_b$, we can write

$$\begin{aligned}
\Pr_S[|\Phi(S)|\geq\epsilon] &= \Pr_S[|R(h_S)-\widehat{R}(h_S)|\geq\epsilon]\\
&= \Pr_S\left[\frac{1}{m}\Big|\sum_{z\in S}\mathop{\mathrm{E}}_Z[c(h_S,Z)]-c(h_S,z)\Big|\geq\epsilon\right]\\
&\leq \Pr_S\left[\frac{1}{m}\Big|\sum_{z\in S_a}\mathop{\mathrm{E}}_Z[c(h_S,Z)]-c(h_S,z)\Big|+\frac{1}{m}\Big|\sum_{z'\in S_b}\mathop{\mathrm{E}}_{Z'}[c(h_S,Z')]-c(h_S,z')\Big|\geq\epsilon\right]\\
&\leq \Pr_S\left[\frac{1}{m}\Big|\sum_{z\in S_a}\mathop{\mathrm{E}}_Z[c(h_S,Z)]-c(h_S,z)\Big|+\frac{\mu bM}{m}\geq\epsilon\right].
\end{aligned}$$

By $\widehat{\beta}$-stability and $\mu a/m\leq 1$, this last term can be bounded as follows

$$\begin{aligned}
&\Pr_S\left[\frac{1}{m}\Big|\sum_{z\in S_a}\mathop{\mathrm{E}}_Z[c(h_S,Z)]-c(h_S,z)\Big|+\frac{\mu bM}{m}\geq\epsilon\right]\leq\\
&\qquad\qquad \Pr_{S_a}\left[\frac{1}{\mu a}\Big|\sum_{z\in S_a}\mathop{\mathrm{E}}_Z[c(h_{S_a},Z)]-c(h_{S_a},z)\Big|+\frac{\mu bM}{m}+2\mu b\widehat{\beta}\geq\epsilon\right].
\end{aligned}$$

The right-hand side can be rewritten in terms of Φ and bounded in terms of a β-mixing coefficient:

$$\begin{aligned}
&\Pr_{S_a}\left[\frac{1}{\mu a}\Big|\sum_{z\in S_a}\mathop{\mathrm{E}}_Z[c(h_{S_a},Z)]-c(h_{S_a},z)\Big|+\frac{\mu bM}{m}+2\mu b\widehat{\beta}\geq\epsilon\right]\\
&\qquad\qquad = \Pr_{S_a}\left[|\Phi(S_a)|+\frac{\mu bM}{m}+2\mu b\widehat{\beta}\geq\epsilon\right]\\
&\qquad\qquad \leq \Pr_{\widetilde{S}_a}\left[|\Phi(\widetilde{S}_a)|+\frac{\mu bM}{m}+2\mu b\widehat{\beta}\geq\epsilon\right]+(\mu-1)\beta(b),
\end{aligned}$$

by applying Lemma 3 to the indicator function of the event $\left\{|\Phi(S_a)| + \frac{\mu b M}{m} + 2\mu b\widehat{\beta} \geq \varepsilon\right\}$. Since $\mathrm{E}_{\widetilde{S}'_a}[|\Phi(\widetilde{S}'_a)|]$ is a constant, the probability in this last term can be rewritten as

$$\begin{aligned}
&\Pr_{\widetilde{S}_a}\left[|\Phi(\widetilde{S}_a)| + \frac{\mu b M}{m} + 2\mu b\widehat{\beta} \geq \varepsilon\right] \\
&\qquad = \Pr_{\widetilde{S}_a}\left[|\Phi(\widetilde{S}_a)| - \mathop{\mathrm{E}}_{\widetilde{S}'_a}[|\Phi(\widetilde{S}'_a)|] + \frac{\mu b M}{m} + 2\mu b\widehat{\beta} \geq \varepsilon - \mathop{\mathrm{E}}_{\widetilde{S}'_a}[|\Phi(\widetilde{S}'_a)|]\right] \\
&\qquad = \Pr_{\widetilde{S}_a}\left[|\Phi(\widetilde{S}_a)| - \mathop{\mathrm{E}}_{\widetilde{S}'_a}[|\Phi(\widetilde{S}'_a)|] \geq \varepsilon'_0\right],
\end{aligned}$$

which ends the proof of the lemma. ■

The last two lemmas will help us prove the main result of this section formulated in the following theorem.

Theorem 19 *Assume a $\widehat{\beta}$-stable algorithm and let ε' denote $\varepsilon - \frac{\mu b M}{m} - 2\mu b\widehat{\beta} - 2a\widehat{\beta}$ as in Lemma 18. Then, for any sample S of size m drawn according to a β-mixing stationary distribution, any choice of the parameters $a, b, \mu > 0$ such that $(a+b)\mu = m$, and $\varepsilon \geq 0$ such that $\varepsilon' \geq 0$, the following generalization bound holds:*

$$\Pr_S\left[|R(h_S) - \widehat{R}(h_S)| \geq \varepsilon\right] \leq \exp\left(\frac{-2\varepsilon'^2 m}{\left(4a\widehat{\beta}m + (a+b)M\right)^2}\right) + (\mu - 1)\beta(b).$$

Proof To prove the statement of theorem, it suffices to bound the probability term appearing in the right-hand side of Equation 18, $\Pr_{\widetilde{S}_a}\left[|\Phi(\widetilde{S}_a)| - \mathrm{E}[|\Phi(\widetilde{S}_a)|] \geq \varepsilon'_0\right]$, which is expressed only in terms of independent blocks. We can therefore apply McDiarmid's inequality by viewing the blocks as i.i.d. "points".

To do so, we must bound the quantity $\left||\Phi(\widetilde{S}_a)| - |\Phi(\widetilde{S}^i_a)|\right|$ where the sequence S_a and S^i_a differ in the ith block. We will bound separately the difference between the generalization errors and empirical errors.[5] The difference in empirical errors can be bounded as follows using the bound on the cost function c:

$$\begin{aligned}
|\widehat{R}(h_{S_a}) - \widehat{R}(h_{S^i_a})| &= \left|\frac{1}{\mu}\left[\sum_{j \neq i} c(h_{S_a}, Z_j) - c(h_{S^i_a}, Z_j)\right] + \frac{1}{\mu}\left[c(h_{S_a}, Z_i) - c(h_{S^i_a}, Z'_i)\right]\right| \\
&\leq 2a\widehat{\beta} + \frac{M}{\mu} = 2a\widehat{\beta} + \frac{(a+b)M}{m}.
\end{aligned}$$

The difference in generalization error can be straightforwardly bounded using $\widehat{\beta}$-stability:

$$|R(h_{S_a}) - R(h_{S^i_a})| = |\mathop{\mathrm{E}}_Z[c(h_{S_a}, Z)] - \mathop{\mathrm{E}}_Z[c(h_{S^i_a}, Z)]| = |\mathop{\mathrm{E}}_Z[c(h_{S_a}, Z) - c(h_{S^i_a}, Z)]| \leq 2a\widehat{\beta}.$$

5. We drop the superscripts on $Z^{(a)}$ since we will not be considering the sequence S_b in what follows.

Using these bounds in conjunction with McDiarmid's inequality yields

$$\Pr_{\widetilde{S}_a}[|\Phi(\widetilde{S}_a)| - \mathop{\mathrm{E}}_{\widetilde{S}'_a}[|\Phi(\widetilde{S}'_a)|] \geq \varepsilon'_0] \leq \exp\left(\frac{-2\varepsilon_0'^2 m}{\left(4a\widehat{\beta}m + (a+b)M\right)^2}\right)$$
$$\leq \exp\left(\frac{-2\varepsilon'^2 m}{\left(4a\widehat{\beta}m + (a+b)M\right)^2}\right).$$

Note that to show the second inequality we make use of Lemma 17 to establish the fact that

$$\varepsilon'_0 = \varepsilon - \frac{\mu bM}{m} - 2\mu b\widehat{\beta} - \mathop{\mathrm{E}}_{\widetilde{S}'_a}[|\Phi(\widetilde{S}'_a)|] \geq \varepsilon - \frac{\mu bM}{m} - 2\mu b\widehat{\beta} - 2a\widehat{\beta} = \varepsilon'.$$

Finally, we make use of Lemma 18 to establish the proof,

$$\Pr_S[|\Phi(S)| \geq \varepsilon] \leq \Pr_{\widetilde{S}_a}\left[|\Phi(\widetilde{S}_a)| - \mathrm{E}[|\Phi(\widetilde{S}_a)|] \geq \varepsilon'_0\right] + (\mu - 1)\beta(b)$$
$$\leq \exp\left(\frac{-2\varepsilon'^2 m}{\left(4a\widehat{\beta}m + (a+b)M\right)^2}\right) + (\mu - 1)\beta(b).$$

This concludes the proof of the theorem. ■

In order to make use of this bound, we must determine the values of parameters b and μ (a is then equal to $\mu/m - u$). There is a trade-off between selecting a large enough value for b to ensure that the mixing term decreases and choosing a large enough value of μ to minimize the remaining terms of the bound. The exact choice of parameters will depend on the type of mixing that is assumed (e.g., algebraic or exponential). In order to choose optimal parameters, it will be useful to view the bound as it holds with high probability, in the following corollary.

Corollary 20 *Assume a $\widehat{\beta}$-stable algorithm and let δ' denote $\delta - (\mu - 1)\beta(b)$. Then, for any sample S of size m drawn according to a β-mixing stationary distribution, any choice of the parameters $a, b, \mu > 0$ such that $(a+b)\mu = m$, and $\delta \geq 0$ such that $\delta' \geq 0$, the following generalization bound holds with probability at least $(1-\delta)$:*

$$|R(h_S) - \widehat{R}(h_S)| < \mu b\left(\frac{M}{m} + 2\widehat{\beta}\right) + 2a\widehat{\beta}\left(4a\widehat{\beta}m + M\frac{m}{\mu}\right)\sqrt{\frac{\log(1/\delta')}{2m}}.$$

In the case of a fast mixing distribution, it is possible to select the values of the parameters to retrieve a bound as in the i.i.d. case, that is, $|R(h_S) - \widehat{R}(h_S)| = O\left(m^{-\frac{1}{2}}\sqrt{\log 1/\delta}\right)$. In particular, for $\beta(b) \equiv 0$, we can choose $a = 0$, $b = 1$, and $\mu = m$ to retrieve the i.i.d. bound of Bousquet and Elisseeff (2001).

In the following, we examine slower mixing algebraic β-mixing distributions, which are thus not close to the i.i.d. scenario. For algebraic mixing, the mixing parameter is defined as $\beta(b) = b^{-r}$. In that case, we wish to minimize the following function in terms of μ and b:

$$s(\mu, b) = \frac{\mu}{b^r} + \frac{m^{3/2}\widehat{\beta}}{\mu} + \frac{m^{1/2}}{\mu} + \mu b\left(\frac{1}{m} + \widehat{\beta}\right). \tag{10}$$

The first term of the function captures the condition $\delta > (\mu+1)\beta(b) \approx \mu/b^r$ and the remaining terms capture the shape of the bound in Corollary 20.

Setting the derivative with respect to each variable μ and b to zero and solving for each parameter results in the following expressions:

$$b = C_r \gamma^{-\frac{1}{r+1}} \qquad \mu = \frac{m^{3/4}\gamma^{\frac{1}{2(r+1)}}}{\sqrt{C_r(1+1/r)}},$$

where $\gamma = (m^{-1}+\widehat{\beta})$ and $C_r = r^{\frac{1}{r+1}}$ is a constant defined by the parameter r.

Now, assuming $\widehat{\beta} = O(m^{-\alpha})$ for some $0 < \alpha \leq 1$, we analyze the convergence behavior of Corollary 20. First, we observe that the terms b and μ have the following asymptotic behavior,

$$b = O\left(m^{\frac{\alpha}{r+1}}\right) \qquad \mu = O\left(m^{\frac{3}{4}-\frac{\alpha}{2(r+1)}}\right).$$

Next, we consider the condition $\delta' > 0$ which is equivalent to,

$$\delta > (\mu-1)\beta(b) = O\left(m^{\frac{3}{4}-\alpha\left(1-\frac{1}{2(r+1)}\right)}\right). \tag{11}$$

In order for the right-hand side of the inequality to converge, it must be the case that $\alpha > \frac{3r+3}{4r+2}$. In particular, if $\alpha = 1$, as is the case for several algorithms in Section 3.5, then it suffices that $r > 1$.

Finally, in order to see how the bound itself converges, we study the asymptotic behavior of the terms of Equation 10 (without the first term, which corresponds to the quantity already analyzed in Equation 11):

$$\underbrace{\frac{m^{3/2}\widehat{\beta}}{\mu} + \mu b\widehat{\beta}}_{(a)} + \underbrace{\frac{m^{1/2}}{\mu} + \frac{\mu b}{m}}_{(b)} = O\left(\underbrace{m^{\frac{3}{4}-\alpha\left(1-\frac{1}{2(r+1)}\right)}}_{(a)} \underbrace{+\, m^{\frac{\alpha}{2(r+1)}-\frac{1}{4}}}_{(b)}\right).$$

This expression can be further simplified by noticing that $(b) \leq (a)$ for all $0 < \alpha \leq 1$ (with equality at $\alpha = 1$). Thus, both the bound and the condition on δ decrease asymptotically as the term in (a), resulting in the following corollary.

Corollary 21 *Assume a $\widehat{\beta}$-stable algorithm with $\widehat{\beta} = O(m^{-1})$ and let $\delta' = \delta - m^{\frac{1}{2(r+1)}-\frac{1}{4}}$. Then, for any sample S of size m drawn according to a algebraic β-mixing stationary distribution, and $\delta \geq 0$ such that $\delta' \geq 0$, the following generalization bound holds with probability at least $(1-\delta)$:*

$$|R(h_S) - \widehat{R}(h_S)| < O\left(m^{\frac{1}{2(r+1)}-\frac{1}{4}}\sqrt{\log(1/\delta')}\right).$$

As in previous bounds $r > 1$ is required for convergence. Furthermore, as expected, a larger mixing parameter r leads to a more favorable bound.

5. Conclusion

We presented stability bounds for both φ-mixing and β-mixing stationary sequences. Our bounds apply to large classes of algorithms, including common algorithms such as SVR, KRR, and SVMs, and extend to non-i.i.d. scenarios existing i.i.d. stability bounds. Since they are algorithm-specific, these bounds can often be tighter than other generalization bounds based on general complexity measures for families of hypotheses. As in the i.i.d. case, weaker notions of stability might help further improve and refine these bounds. These stability bounds complement general data-dependent learning bounds we have shown elsewhere for stationary β-mixing sequences using the notion of Rademacher complexity (Mohri and Rostamizadeh, 2009).

The stability bounds we presented can be used to analyze the properties of stable algorithms when used in the non-i.i.d settings studied. But, more importantly, they can serve as a tool for the design of novel and accurate learning algorithms. Of course, some mixing properties of the distributions need to be known to take advantage of the information supplied by our generalization bounds. In some problems, it is possible to estimate the shape of the mixing coefficients. This should help devising such algorithms.

Acknowledgments

We thank the editor and the reviewers for several comments that helped improve the original version of this paper.

References

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

Olivier Bousquet and André Elisseeff. Algorithmic stability and generalization performance. In *Advances in Neural Information Processing Systems*, 2001.

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

Jean-René Chazottes, Pierre Collet, Christof Külske, and Frank Redig. Concentration inequalities for random fields via coupling. *Probability Theory and Related Fields*, 137(1):201–225, 2007.

Corinna Cortes and Vladimir N. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

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

Paul Doukhan. *Mixing: Properties and Examples*. Springer-Verlag, 1994.

Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Computational Learning Theory*, pages 152–162, 1997.

Leo Kontorovich. *Measure Concentration of Strongly Mixing Processes with Applications*. PhD thesis, Carnegie Mellon University, 2007.

Leo Kontorovich and Kavita Ramanan. Concentration inequalities for dependent random variables via the martingale method. *Annals of Probability*, 36(6):2126–2158, 2008.

Aurélie Lozano, Sanjeev Kulkarni, and Robert Schapire. Convergence and consistency of regularized boosting algorithms with stationary β-mixing observations. In *Advances in Neural Information Processing Systems*, 2006.

Katalin Marton. Measure concentration for a class of random processes. *Probability Theory and Related Fields*, 110(3):427–439, 1998.

Davide Mattera and Simon 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, USA, 1999. ISBN 0-262-19416-3.

Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press, 1989.

Ron Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39(1):5–34, April 2000.

Dharmendra Modha and Elias Masry. On the consistency in nonparametric estimation under mixing assumptions. *IEEE Transactions of Information Theory*, 44:117–133, 1998.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for non-iid processes. In *Advances in Neural Information Processing Systems*, 2007.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In *Advances in Neural Information Processing Systems (NIPS 2008)*, pages 1097–1104, Vancouver, Canada, 2009. MIT Press.

Klaus-Robert Müller, Alex Smola, Gunnar Rätsch, Bernhard Schölkopf, Jens Kohlmorgen, and Vladimir Vapnik. Predicting time series with support vector machines. In *Proceedings of the International Conference on Artificial Neural Networks*, Lecture Notes in Computer Science, pages 999–1004. Springer, 1997.

Paul-Marie Samson. Concentration of measure inequalities for Markov chains and-mixing processes. *Annals Probability*, 28(1):416–461, 2000.

Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 515–521. Morgan Kaufmann Publishers Inc., 1998.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

Mathukumalli Vidyasagar. *Learning and Generalization: with Applications to Neural Networks*. Springer, 2003.

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

Shuheng Zhou, John Lafferty, and Larry Wasserman. Time varying undirected graphs. In *Proceedings of the 21st Annual Conference on Learning Theory*, 2008.