Generalization Bounds for Time Series Prediction with Non-stationary Processes

Vitaly Kuznetsov[1] and Mehryar Mohri[1,2]

[1] Courant Institute of Mathematical Sciences,
251 Mercer street, New York, NY 10012, USA
[2] Google Research, 111 8th Avenue, New York, NY 10012, USA
{vitaly,mohri}@cims.nyu.edu

Abstract. This paper presents the first generalization bounds for time series prediction with a non-stationary mixing stochastic process. We prove Rademacher complexity learning bounds for both average-path generalization with non-stationary β-mixing processes and path-dependent generalization with non-stationary ϕ-mixing processes. Our guarantees are expressed in terms of β- or ϕ-mixing coefficients and a natural measure of discrepancy between training and target distributions. They admit as special cases previous Rademacher complexity bounds for non-i.i.d. stationary distributions, for independent but not identically distributed random variables, or for the i.i.d. case. We show that, using a new sub-sample selection technique we introduce, our bounds can be tightened under the natural assumption of convergent stochastic processes. We also prove that fast learning rates can be achieved by extending existing local Rademacher complexity analysis to non-i.i.d. setting.

Keywords: Generalization bounds, time series, mixing, stationary processes, fast rates, local Rademacher complexity.

1 Introduction

Given a sample $((X_1, Y_1), \ldots, (X_m, Y_m))$ of pairs in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, the standard supervised learning task consists of selecting, out of a class of functions H, a hypothesis $h\colon \mathcal{X} \to \mathcal{Y}$ that admits a small expected loss measured using some specified loss function $L\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. The common assumption in the statistical learning theory and the design of algorithms is that samples are drawn i.i.d. from some unknown distribution and generalization in this scenario has been extensively studied in the past. However, for many problems such as time series prediction, the i.i.d. assumption is too restrictive and it is important to analyze generalization in the absence of that condition. A variety of relaxations of this i.i.d. setting have been proposed in the machine learning and statistics literature. In particular, the scenario in which observations are drawn from a stationary mixing distribution has become standard and has been adopted by most previous studies [1, 10–12, 18, 20]. In this work, we seek to analyze generalization under the

more realistic assumption of non-stationary data. This covers a wide spectrum of stochastic processes considered in applications, including Markov chains, which are non-stationary.

Suppose we are given a doubly infinite sequence of $\mathcal{Z}$-valued random variables $\{Z_t\}_{t=-\infty}^{\infty}$ jointly distributed according to $\mathbf{P}$. We will write $\mathbf{Z}_a^b$ to denote a vector $(Z_a, Z_{a+1}, \ldots, Z_b)$ where a and b are allowed to take values $-\infty$ and ∞. Similarly, $\mathbf{P}_a^b$ denotes the distribution of $\mathbf{Z}_a^b$. Following [4], we define β-mixing coefficients for $\mathbf{P}$ as follows. For each positive integer a, we set

$$\beta(a) = \sup_t \|\mathbf{P}_{-\infty}^t \otimes \mathbf{P}_{t+a}^{\infty} - \mathbf{P}_{-\infty}^t \wedge \mathbf{P}_{t+a}^{\infty}\|_{TV}, \tag{1}$$

where $\mathbf{P}_{-\infty}^t \wedge \mathbf{P}_{t+a}^{\infty}$ denotes the joint distribution of $\mathbf{Z}_{-\infty}^t$ and $\mathbf{Z}_{t+a}^{\infty}$. Recall that the total variation distance $\|\cdot\|_{TV}$ between two probability measures P and Q defined on the same σ-algebra of events $\mathcal{G}$ is given by $\|P - Q\|_{TV} = \sup_{A \in \mathcal{G}} |P(A) - Q(A)|$. We say that $\mathbf{P}$ is β-mixing (or absolutely regular) if $\beta(a) \to 0$ as $a \to \infty$. Roughly speaking, this means that the future has a sufficiently weak dependence on the distant past. We remark that β-mixing coefficients can be defined equivalently as follows:

$$\beta(a) = \sup_t \mathbb{E}_{\mathbf{Z}_{-\infty}^t} \Big[\|\mathbf{P}_{t+a}^{\infty}(\cdot|\mathbf{Z}_{-\infty}^t) - \mathbf{P}_{t+a}^{\infty}\|_{TV} \Big], \tag{2}$$

where $\mathbf{P}(\cdot|\cdot)$ denotes conditional probability measure [4]. Another standard measure of the dependence of the future on the past is the φ-mixing coefficient defined for any $a > 0$ by

$$\varphi(a) = \sup_t \sup_{B \in \mathcal{F}_t} \|\mathbf{P}_{t+a}^{\infty}(\cdot|B) - \mathbf{P}_{t+a}^{\infty}\|_{TV}, \tag{3}$$

where $\mathcal{F}_t$ is the σ-algebra generated by $\mathbf{Z}_{-\infty}^t$. A distribution $\mathbf{P}$ is said to be φ-mixing if $\varphi(a) \to 0$ as $a \to \infty$. Note that $\beta(a) \leq \varphi(a)$, so any φ-mixing distribution is necessarily β-mixing. We also recall that a sequence of random variables $\mathbf{Z}_{-\infty}^{\infty}$ is (strictly) stationary provided that, for any t and any non-negative integers m and k, $\mathbf{Z}_t^{t+m}$ and $\mathbf{Z}_{t+k}^{t+m+k}$ have the same distribution.

Unlike the i.i.d. case where $\mathbb{E}[L(h(X), Y)]$ is used to measure the generalization error of h, in the case of time series prediction, there is no unique measure commonly used to assess the quality of a given hypothesis h. One approach consists of seeking a hypothesis h that performs well in the near future, given the observed trajectory of the process. That is, we would like to achieve a small *path-dependent* generalization error

$$\mathcal{L}_{T+s}(h) = \mathbb{E}_{Z_{T+s}}[L(h(X_{T+s}), Y_{T+s})|\mathbf{Z}_1^T], \tag{4}$$

where $s \geq 1$ is fixed. To simplify the notation, we will often write $\ell(h, z) = L(h(x), y)$, where $z = (x, y)$. For time series prediction tasks, we often receive a sample $\mathbf{Y}_1^T$ and wish to forecast Y_{T+s}. A large class of (bounded-memory) auto-regressive models uses q past observations $\mathbf{Y}_{T-q+1}^T$ to predict Y_{T+s}. Our scenario includes this setting as a special case where we take $\mathcal{X} = \mathcal{Y}^q$ and

$Z_{t+s} = (\mathbf{Y}_{t-q+1}^{t}, Y_{t+s})$.[3] The generalization ability of stable algorithms with error defined by (4) was studied by Mohri and Rostamizadeh [12].

Alternatively, one may wish to perform well in the near future when being on some "average" trajectory. This leads to the *averaged* generalization error:

$$\bar{\mathcal{L}}_{T+s}(h) = \mathbb{E}_{\mathbf{Z}_1^T}[\mathcal{L}_{T+s}(h)] = \mathbb{E}_{Z_{T+s}}[\ell(h, Z_{T+s})]. \tag{5}$$

We note that $\bar{\mathcal{L}}_{T+s}(h) = \mathcal{L}_{T+s}(h)$ when the training and testing sets are independent. The pioneering work of Yu [20] led to VC-dimension bounds for $\bar{\mathcal{L}}_{T+s}$ under the assumption of stationarity and β-mixing. Later, Meir [10] used that to derive generalization bounds in terms of covering numbers of H. These results have been further extended by Mohri and Rostamizadeh [11] to data-dependent learning bounds in terms of the Rademacher complexity of H.

Most of the generalization bounds for non-i.i.d. scenarios that can be found in the machine learning and statistics literature assume that observations come from a (strictly) stationary distribution. The only exception that we are aware of is the work of Agarwal and Duchi [1], who present bounds for stable on-line learning algorithms under the assumptions of suitably convergent process.[4] The main contribution of our work is the first generalization bounds for both $\mathcal{L}_{T+s}$ and $\bar{\mathcal{L}}_{T+s}$ when the data is generated by a non-stationary mixing stochastic process. These results provide a sufficient condition for the predictive PAC learnability of Pestov [3, 14]. Next, we strengthen our assumptions and give generalization bounds for convergent processes. In doing so, we establish sufficient conditions for the predictive PAC learnability of Shalizi and Kontorovich [17]. These results are algorithm-agnostic analogues of the algorithm-dependent bounds of Agarwal and Duchi [1]. In [1], Agarwal and Duchi also prove fast convergence rates when a strongly convex loss is used. Similarly, Steinwart and Christmann [18] showed that regularized learning algorithms admit faster convergence rates under the assumptions of mixing and stationarity. We conclude this paper by showing that this is in fact a general phenomenon. We use local Rademacher complexity techniques [2] to establish faster convergence rates for stationary or convergent mixing processes.

A key ingredient of the bounds we present is the notion of *discrepancy* between two probability distributions that was used by Mohri and Muñoz Medina [13] to give generalization bounds for sequences of independent (but not identically distributed) random variables. In our setting, discrepancy can be defined as

$$d(t_1, t_2) = \sup_{h \in H} |\mathcal{L}_{t_1}(h) - \mathcal{L}_{t_2}(h)| \tag{6}$$

and similarly we can define $\bar{d}(t_1, t_2)$, where we replace $\mathcal{L}_t$ with $\bar{\mathcal{L}}_t$. Discrepancy is a natural measure of the non-stationarity of a stochastic process with respect to

[3] Observe that if $\mathbf{Y}$ is β-mixing, then so is $\mathbf{Z}$ and $\beta_{\mathbf{Z}}(a) = \beta_{\mathbf{Y}}(a - q)$. Similarly, the φ-mixing assumption is also preserved. It is an open problem (posed by Meir [10]) to derive generalization bounds for unbounded-memory models.

[4] Agarwal and Duchi [1] additionally assume that distributions are absolutely continuous and that the loss function is convex and Lipschitz.

the hypothesis class H and a loss function L. For instance, if the process is strictly stationary then $\bar{d}(t_1, t_2) = 0$ for all $t_1, t_2 \in \mathbb{Z}$. As a more interesting example, consider a weakly stationary stochastic process,[5] together with a squared loss L and a set of linear hypothesis $H = \{\mathbf{Y}_{t-q+1}^T \mapsto w \cdot \mathbf{Y}_{t-q+1}^T \colon w \in \mathbb{R}^q\}$. It can be shown that in this case we again have $\bar{d}(t_1, t_2) = 0$ for all $t_1, t_2 \in \mathbb{Z}$. An additional advantage of the discrepancy measure is that it can be replaced by an upper bound that, under mild conditions, can be estimated from data [8, 6].

The rest of this paper is organized as follows. In Section 2 we discuss the main technical tool used to derive our bounds. Section 3 and Section 4 present learning guarantees for averaged and path-dependent errors respectively. In Section 5 we analyze generalization with convergent processes. We conclude with fast learning rates for the non-i.i.d. setting in Section 6.

2 Independent Blocks and Sub-sample Selection

The first step towards our generalization bounds is to reduce the setting of a mixing stochastic process to a simpler scenario of a sequence of independent random variables, where we can take advantage of the known concentration results. One way to achieve this is via the independent block technique introduced by Yu [20] which we now describe.

We can divide a given sample $\mathbf{Z}_1^T$ into $2m$ blocks such that each block has size a_i and we require $T = \sum_{i=1}^{2m} a_i$. In other words, we consider a sequence of random vectors $\mathbf{Z}(i) = \mathbf{Z}_{l(i)}^{u(i)}, i = 1, \ldots, 2m$ where $l(i) = 1 + \sum_{j=1}^{i-1} a_j$ and $u(i) = \sum_{j=1}^{i} a_j$. It will be convenient to refer to even and odd blocks separately. We will write $\mathbf{Z}^o = (\mathbf{Z}(1), \mathbf{Z}(3) \ldots, \mathbf{Z}(2m-1))$ and $\mathbf{Z}^e = (\mathbf{Z}(2), \mathbf{Z}(4), \ldots, \mathbf{Z}(2m))$. In fact, we will often work with blocks that are independent.

Let $\widetilde{\mathbf{Z}}^o = (\widetilde{\mathbf{Z}}(1), \ldots, \widetilde{\mathbf{Z}}(2m-1))$ where $\widetilde{\mathbf{Z}}(i)$, $i = 1, 3, \ldots, 2m-1$, are independent and each $\widetilde{\mathbf{Z}}(i)$ has the same distribution as $\mathbf{Z}(i)$. We construct $\widetilde{\mathbf{Z}}^e$ in the same way. The following result due to Yu [20] enables us to relate sequences of dependent and independent blocks.

Proposition 1. *Let g be a real-valued Borel measurable function such that $-M_1 \leq g \leq M_2$ for some $M_1, M_2 \geq 0$. Then, the following holds:*

$$|\mathbb{E}[g(\widetilde{\mathbf{Z}}^o)] - \mathbb{E}[g(\mathbf{Z}^o)]| \leq (M_1 + M_2) \sum_{i=1}^{m-1} \beta(a_{2i}).$$

The proof of this result is given in [20], which in turn is based on [5] and [19]. We present a sketch of the main steps of the proof as these will be useful for us as stand-alone results.

[5] A process $\mathbf{Z}$ is weakly stationary if $\mathbb{E}[Z_t]$ is a constant function of t and $\mathbb{E}[Z_{t_1} Z_{t_2}]$ only depends on $t_1 - t_2$.

Lemma 1. *Let Q and P be probability measures on $(\Omega, \mathcal{F})$ and let $h\colon \Omega \to \mathbb{R}$ be a Borel measurable function such that $-M_1 \leq h \leq M_2$ for some $M_1, M_2 \geq 0$. Then*

$$|\mathbb{E}_Q[h] - \mathbb{E}_P[h]| \leq (M_1 + M_2)\|P - Q\|_{TV}.$$

The proof of Lemma 1 can be found in [5, 19, 20]. Lemma 1 extended via induction yields the following result. See [5] for further details.

Lemma 2. *Let $m \geq 1$ and $(\prod_{k=1}^m \Omega_k, \prod_{k=1}^m \mathcal{F}_k)$ be a measure space with P a measure on this space and P_j the marginal on $(\prod_{k=1}^j \Omega_k, \prod_{k=1}^j \mathcal{F}_k)$. Let Q_j be a measure on $(\Omega_j, \mathcal{F}_j)$ and define*

$$\beta_j = \mathbb{E}\left[\left\| P_{j+1}\Big(\cdot \,|\, \prod_{k=1}^{j} \mathcal{F}_k \Big) - Q_{j+1} \right\|_{TV}\right].$$

Then, for any Borel measurable function $h\colon \prod_{k=1}^m \Omega_k \to \mathbb{R}$ such that $-M_1 \leq h \leq M_2$ for some $M_1, M_2 \geq 0$, the following holds

$$|\mathbb{E}_P[h] - \mathbb{E}_Q[h]| \leq (M_1 + M_2) \sum_{j=1}^{m-1} \beta_j$$

where $Q = Q_1 \otimes Q_2 \otimes \ldots \otimes Q_m$.

Proposition 1 now follows from Lemma 2 by taking Q_j to be the marginal of P on $(\Omega_j, \mathcal{F}_j)$ and applying it to the case of independent blocks.

Proposition 1 is not the only way to relate mixing and independent cases. Next, we present another technique that we term *sub-sample selection*, which is particularly useful when the process is convergent. Suppose we are given a sample $\mathbf{Z}_1^T$. Fix $a \geq 1$ such that $T = ma$ for some $m \geq 1$ and define a sub-sample $\mathbf{Z}^{(j)} = (Z_{1+j}, \ldots, Z_{m-1+j})$, $j = 0, \ldots, a-1$. An application of Lemma 2 yields the following result.

Proposition 2. *Let g be a real-valued Borel measurable function such that $-M_1 \leq g \leq M_2$ for some $M_1, M_2 \geq 0$. Then*

$$|\mathbb{E}[g(\widetilde{\mathbf{Z}}_{\Pi})] - \mathbb{E}[g(\mathbf{Z}^{(j)})]| \leq (M_1 + M_2)(m-1)\beta(a),$$

where $\widetilde{\mathbf{Z}}_{\Pi}$ is an i.i.d. sample of size m from a distribution Π and $\beta(a) = \sup_t \mathbb{E}[\|\mathbb{P}_{t+a}(\cdot|\mathbf{Z}_1^t) - \Pi\|_{TV}]$.

Proposition 2 is commonly applied with Π being the stationary probability measure of a convergent process.

3 Generalization Bound for the Averaged Error

In this section, we derive a generalization bound for averaged error $\bar{\mathcal{L}}_{T+s}$. Given a sample $\mathbf{Z}_1^T$ generated by a (β-)mixing process,[6] we define $\Phi(\mathbf{Z}_1^T)$ as follows:

$$\Phi(\mathbf{Z}_1^T) = \sup_{h \in H} \left(\bar{\mathcal{L}}_{T+s}(h) - \frac{1}{T} \sum_{t=1}^{T} \ell(h, Z_t) \right). \tag{7}$$

We will also use I_1 to denote the set of indices of the elements from the sample $\mathbf{Z}_1^T$ that are contained in the odd blocks. Similarly, I_2 is used for elements in the even blocks.

We establish our bounds in a series of lemmas. We start by proving a concentration result for dependent non-stationary data.

Lemma 3. *Let L be a loss function bounded by M and H an arbitrary hypothesis set. For any $a_1, \ldots, a_{2m} > 0$ such that $T = \sum_{i=1}^{2m} a_i$, partition the given sample $\mathbf{Z}_1^T$ into blocks as described in Section 2. Then, for any $\epsilon > \max(\mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^o)], \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^e)])$, the following holds:*

$$\mathbb{P}(\Phi(\mathbf{Z}_1^T) > \epsilon) \leq \mathbb{P}(\Phi(\widetilde{\mathbf{Z}}^o) - \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^o)] > \epsilon_1) + \mathbb{P}(\Phi(\widetilde{\mathbf{Z}}^e) - \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^e)] > \epsilon_2) + \sum_{i=2}^{m-1} \beta(a_i),$$

where $\epsilon_1 = \epsilon - \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^o)]$ and $\epsilon_2 = \epsilon - \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^e)]$.

Proof. By convexity of the supremum $\Phi(\mathbf{Z}_1^T) \leq \frac{|I_1|}{T}\Phi(\mathbf{Z}^o) + \frac{|I_2|}{T}\Phi(\mathbf{Z}^e)$. Since $|I_1| + |I_2| = T$, for $\frac{|I_1|}{T}\Phi(\mathbf{Z}^o) + \frac{|I_2|}{T}\Phi(\mathbf{Z}^e)$ to exceed ϵ at least one element of $\{\Phi(\mathbf{Z}^o), \Phi(\mathbf{Z}^e)\}$ must be greater than ϵ. Thus, by the union bound, we can write

$$\begin{aligned}\mathbb{P}(\Phi(\mathbf{Z}_1^T) > \epsilon) &\leq \mathbb{P}(\Phi(\mathbf{Z}^o) > \epsilon) + \mathbb{P}(\Phi(\mathbf{Z}^e) > \epsilon) \\ &= \mathbb{P}(\Phi(\mathbf{Z}^o) - \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^o)] > \epsilon_1) + \mathbb{P}(\Phi(\mathbf{Z}^e) - \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^e)] > \epsilon_2).\end{aligned}$$

We apply Proposition 1 to the indicator functions of the events $\{\Phi(\mathbf{Z}^o) - \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^o)] > \epsilon_1\}$ and $\{\Phi(\mathbf{Z}^e) - \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^e)] > \epsilon_2\}$ to complete the proof. □

Lemma 4. *Under the same assumptions as in Lemma 3, the following holds:*

$$\mathbb{P}(\Phi(\mathbf{Z}_1^T) > \epsilon) \leq \exp\left(\frac{-2T^2\epsilon_1^2}{\|\mathbf{a}^o\|_2^2 M^2}\right) + \exp\left(\frac{-2T^2\epsilon_2^2}{\|\mathbf{a}^e\|_2^2 M^2}\right) + \sum_{i=2}^{m-1} \beta(a_i),$$

where $\mathbf{a}^o = (a_1, a_3, \ldots, a_{2m-1})$ and $\mathbf{a}^e = (a_2, a_4, \ldots, a_{2m})$.

[6] All the results of this section hold for a slightly weaker notion of β-mixing with $\beta(a) = \sup_t \mathbb{E}\|\mathbf{P}_{t+a}(\cdot|\mathbf{Z}_{-\infty}^t) - \mathbf{P}_{t+a}\|_{TV}$.

Proof. We apply McDiarmid's inequality [9] to the sequence of independent blocks. We note that if $\widetilde{\mathbf{Z}}^o$ and $\widetilde{\mathbf{Z}}$ are two sequences of independent (odd) blocks that differ only by one block (say block i) then $\Phi(\widetilde{\mathbf{Z}}^o) - \Phi(\widetilde{\mathbf{Z}}) \leq a_i \frac{M}{T}$ and it follows from McDiarmid's inequality that

$$\mathbb{P}(\Phi(\widetilde{\mathbf{Z}}^o) - \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^o)] > \epsilon_1) \leq \exp\left(\frac{-2T^2\epsilon_1^2}{\|\mathbf{a}^o\|_2^2 M^2}\right).$$

Using the same argument for $\widetilde{\mathbf{Z}}^e$ finishes the proof of this lemma. □

The next step is to bound $\max(\mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^o)], \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^e)])$. The bound that we give is in terms of *block* Rademacher complexity defined by

$$\mathfrak{R}(\widetilde{\mathbf{Z}}^o) = \frac{1}{|I_1|}\mathbb{E}\Big[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i\, l\big(h, \mathbf{Z}(2i-1)\big)\Big], \tag{8}$$

where σ_i is a sequence of Rademacher random variables and $l(h, \mathbf{Z}(2i-1)) = \sum_{t\in I_1\cap\mathbf{Z}(2i-1)} \ell(h, Z_t)$. Below we will show that if the block size is constant (i.e. $a_i = a$), then the block complexity can be bounded in terms of the regular Rademacher complexity.

Lemma 5. *For $j = 1, 2$, let $\Delta^j = \frac{1}{|I_j|}\sum_{t\in I_j}\bar{d}(t, T+s)$, which is an average discrepancy. Then, the following bound holds:*

$$\max(\mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^o)], \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^e)]) \leq 2\max(\mathfrak{R}(\widetilde{\mathbf{Z}}^o), \mathfrak{R}(\widetilde{\mathbf{Z}}^e)) + \max(\Delta^1, \Delta^2). \tag{9}$$

Proof. In the course of this proof Z_t, denotes a sample drawn according to the distribution of $\widetilde{\mathbf{Z}}^o$ (and not that of $\mathbf{Z}^o$). Using the sub-additivity of the supremum and the linearity of expectation, we can write

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{h\in H}\bar{\mathcal{L}}_{T+s}(h) - \frac{1}{|I_1|}\sum_{t\in I_1}\ell(h, Z_t)\Big] \\
&= \mathbb{E}\Big[\sup_{h\in H}\bar{\mathcal{L}}_{T+s}(h) - \frac{1}{|I_1|}\sum_{t\in I_1}\bar{\mathcal{L}}_t(h) + \frac{1}{|I_1|}\sum_{t\in I_1}\bar{\mathcal{L}}_t(h) - \frac{1}{|I_1|}\sum_{t\in I_1}\ell(h, Z_t)\Big] \\
&\leq \mathbb{E}\Big[\sup_{h\in H}\bar{\mathcal{L}}_{T+s}(h) - \frac{1}{|I_1|}\sum_{t\in I_1}\bar{\mathcal{L}}_t(h) + \sup_{h\in H}\frac{1}{|I_1|}\sum_{t\in I_1}\bar{\mathcal{L}}_t(h) - \frac{1}{|I_1|}\sum_{t\in I_1}\ell(h, Z_t)\Big] \\
&= \frac{1}{|I_1|}\sum_{t\in I_1}\sup_{h\in H}\big|\bar{\mathcal{L}}_{T+s}(h) - \bar{\mathcal{L}}_t(h)\big| + \frac{1}{|I_1|}\mathbb{E}\Big[\sup_{h\in H}\sum_{t\in I_1}\bar{\mathcal{L}}_t(h) - \sum_{t\in I_1}\ell(h, Z_t)\Big] \\
&= \Delta^1 + \frac{1}{|I_1|}\mathbb{E}\Big[\sup_{h\in H}\sum_{i=1}^{m}\mathbb{E}[l(h, \widetilde{\mathbf{Z}}(2i-1))] - l(h, \widetilde{\mathbf{Z}}(2i-1))\Big].
\end{aligned}$$

The second term can be written as

$$A = \frac{1}{|I_1|}\mathbb{E}\Big[\sup_{h\in H}\sum_{i=1}^{m}A_i(h)\Big],$$

with $A_i(h) = \mathbb{E}[l(h, \widetilde{\mathbf{Z}}(2i-1))] - l(h, \widetilde{\mathbf{Z}}(2i-1))$ for all $i \in [1, m]$. Since the terms $A_i(h)$ are all independent, the same proof as that of the standard i.i.d. symmetrization bound in terms of the Rademacher complexity applies and A can be bounded by $\mathfrak{R}(\widetilde{\mathbf{Z}}^o)$. Using the same arguments for even blocks completes the proof. □

Combining Lemma 4 and Lemma 5 leads directly to the main result of this section.

Theorem 1. *With the assumptions of Lemma 3, for any $\delta > \sum_{i=2}^{m-1} \beta(a_i)$, with probability $1-\delta$, the following holds for all hypotheses $h \in H$:*

$$\bar{\mathcal{L}}_{T+s}(h) \leq \frac{1}{T}\sum_{t=1}^{T} \ell(h, Z_t) + 2\max(\mathfrak{R}(\widetilde{\mathbf{Z}}^o), \mathfrak{R}(\widetilde{\mathbf{Z}}^e)) + \max(\Delta^1, \Delta^2) + M\max(\|\mathbf{a}^e\|_2, \|\mathbf{a}^e\|_2)\sqrt{\frac{\log \frac{2}{\delta'}}{2T^2}},$$

where $\delta' = \delta - \sum_{i=2}^{m-1} \beta(a_i)$.

The learning bound of Theorem 1 indicates the challenges faced by the learner when presented with data drawn from a non-stationary stochastic process. In particular, the presence of the term $\max(\Delta^1, \Delta^2)$ in the bound shows that generalization in this setting depends on the "degree" of non-stationarity of the underlying process. The dependency in the training instances reduces the effective size of the sample from T to $(T/(\|\mathbf{a}^e\|_2 + \|\mathbf{a}^e\|_2))^2$. Observe that for a general non-stationary process the learning bounds presented may not converge to zero as a function of the sample size, due to the discrepancies between the training and target distributions. In Section 5 and Section 6, we will describe some natural assumptions under which this convergence does occur.

When the same size a is used for all the blocks considered in the analysis, thus $T = 2ma$, then the block Rademacher complexity terms can be replaced with standard Rademacher complexities. Indeed, in that case, we can group the summands in the definition of the block complexity according to subsamples $\mathbf{Z}^{(j)}$ and use the sub-additivity of the supremum to find that $\mathfrak{R}(\widetilde{\mathbf{Z}}^o) \leq \frac{1}{a}\sum_{j=1}^{a} \mathfrak{R}_m(\widetilde{\mathbf{Z}}^{(j)})$, where $\mathfrak{R}_m(\widetilde{\mathbf{Z}}^{(j)}) = \frac{1}{m}\mathbb{E}[\sup_{h\in H}\sum_{i=1}\sigma_i \ell(h, Z_{i,j})]$ with $(\sigma_i)_i$ a sequence of Rademacher random variables and $(Z_{i,j})_{i,j}$ a sequence of independent random variables such that $Z_{i,j}$ is distributed according to the law of $Z_{a(2i-1)+j}$ from $\mathbf{Z}_1^T$. This leads to the following perhaps more informative but somewhat less tight bound:

$$\bar{\mathcal{L}}_{T+s}(h) \leq \frac{1}{T}\sum_{t=1}^{T} \ell(h, Z_t) + \frac{2}{a}\sum_{j=1}^{2a} \mathfrak{R}_m(\mathbf{Z}^{(j)}) + \frac{2}{T}\sum_{t=1}^{T} \bar{d}(t, T+s) + M\sqrt{\frac{\log \frac{2}{\delta'}}{8m}}.$$

If the process is stationary, then we recover as a special case the generalization bound of [11]. If $\mathbf{Z}_1^T$ is a sequence of independent but not identically distributed random variables, we recover the results of [13]. In the i.i.d. case, Theorem 1 reduces to the generalization bounds of Koltchinskii and Panchenko [7].

4 Generalization Bound for the Path-Dependent Error

In this section we give generalization bounds for a path-dependent error $\mathcal{L}_{T+s}$ under the assumption that the data is generated by a (φ-)mixing non-stationary process.[7] In this section, we will use $\Phi(\mathbf{Z}_1^T)$ to denote the same quantity as in (7) except that $\bar{\mathcal{L}}_{T+s}$ is replaced with $\mathcal{L}_{T+s}$.

The key technical tool that we will use is the version of McDiarmid's inequality for dependent random variables, which requires a bound on the differences of conditional expectations of Φ (see Corollary 6.10 in [9]). We start with the following adaptation of Lemma 1 to this setting.

Lemma 6. *Let $\mathbf{Z}_1^T$ be a sequence of $\mathcal{Z}$-valued random variables and suppose that $g\colon \mathcal{Z}^{k+j} \to \mathbb{R}$ is a Borel-measurable function such that $-M_1 \leq g \leq M_2$ for some $M_1, M_2 \geq 0$. Then, for any $z_1, \ldots, z_k \in \mathcal{Z}$, the following bound holds:*

$$|\mathbb{E}[g(Z_1, \ldots, Z_k, Z_{T-j+1}, \ldots, Z_T)|z_1, \ldots, z_k] - \mathbb{E}[g(z_1, \ldots, z_k, Z_{T-j+1}, \ldots, Z_T)]| \leq (M_1 + M_2)\varphi(T + 1 - (k + j)).$$

Proof. This result follows from an application of Lemma 1:

$$\begin{aligned}|\mathbb{E}[g(Z_1, \ldots, Z_k, Z_{T-j+1}, \ldots, Z_T)|z_1, \ldots, z_k] - \mathbb{E}[g(z_1, \ldots, z_k, Z_{T-j+1}, \ldots, Z_T)]| \\ \leq (M_1 + M_2)\|\mathbf{P}_{T-j+1}^T(\cdot|z_1, \ldots, z_k) - \mathbf{P}_{T-j+1}^T\|_{TV} \\ \leq (M_1 + M_2)\varphi(T + 1 - (k + j)),\end{aligned}$$

where the second inequality follows from the definition of φ-mixing coefficients. □

Lemma 7. *For any $z_1, \ldots, z_k, z_k' \in \mathcal{Z}$ and any $0 \leq j \leq T - k$ with $k > 1$, the following holds:*

$$\left|\mathbb{E}[\Phi(\mathbf{Z}_1^T)|z_1, \ldots, z_k] - \mathbb{E}[\Phi(\mathbf{Z}_1^T)|z_1, \ldots, z_k']\right| \leq 2M\left(\tfrac{j+1}{T} + \gamma\varphi(j + 2) + \varphi(s)\right),$$

where $\gamma = 1$ iff $j + k < T$ and 0 otherwise. Moreover, if $\mathcal{L}_{T+s}(h) = \bar{\mathcal{L}}_{T+s}(h)$, then the term $\varphi(s)$ can be omitted from the bound.

Proof. First, we observe that using Lemma 6 we have $|\mathcal{L}_{T+s}(h) - \bar{\mathcal{L}}_{T+s}(h)| \leq M\varphi(s)$. Next, we use this result, the properties of conditional expectation and Lemma 6 to show that $\mathbb{E}[\Phi(\mathbf{Z}_1^T)|z_1, \ldots, z_k]$ is bounded by

$$\begin{aligned}&\mathbb{E}\left[\sup_{h\in H}\left(\bar{\mathcal{L}}_{T+s}(h) - \frac{1}{T}\sum_{t=1}^{T}\ell(h, Z_t)\right)\middle| z_1, \ldots, z_k\right] + M\varphi(s) \\ &\leq \mathbb{E}\left[\sup_{h\in H}\left(\bar{\mathcal{L}}_{T+s}(h) - \frac{1}{T}\sum_{t=k+j}^{T}\ell(h, Z_t) - \frac{1}{T}\sum_{t=1}^{k-1}\ell(h, Z_t)\right)\middle| z_1, \ldots, z_k\right] + \eta \\ &\leq \mathbb{E}\left[\sup_{h\in H}\left(\bar{\mathcal{L}}_{T+s}(h) - \frac{1}{T}\sum_{t=k+j}^{T}\ell(h, Z_t) - \frac{1}{T}\sum_{t=1}^{k-1}\ell(h, z_t)\right)\right] + M\gamma\varphi(j + 2) + \eta,\end{aligned}$$

[7] As in Section 3, we can weaken the notion of φ-mixing by using $\varphi(a) = \sup_t \sup_{B\in\mathcal{F}_t} \|\mathbf{P}_{t+a}(\cdot|B) - \mathbf{P}_{t+a}\|_{TV}$.

where $\eta = M(\frac{j}{T} + \varphi(s))$. Using a similar argument to bound $\mathbb{E}[\Phi(\mathbf{Z}_1^T)|z_1, \ldots, z'_k]$ from below by $-M(\gamma\varphi(j+2) + \frac{j}{T} + \varphi(s))$ and taking the difference completes the proof. □

The last ingredient that we will need to establish a generalization bound for $\mathcal{L}_{T+s}$ is a bound on $\mathbb{E}[\Phi]$. The bound we present is in terms of a discrepancy measure and the sequential Rademacher complexity introduced in [15].

Lemma 8. *The following bound holds*

$$\mathbb{E}[\Phi(\mathbf{Z}_1^T)] \leq \mathbb{E}[\Delta] + 2\mathfrak{R}^{seq}_{T-s}(H_\ell) + M\frac{s-1}{T},$$

where $\mathfrak{R}^{seq}_{T-s}(H_\ell)$ *is the sequential Rademacher complexity of the function class* $H_\ell = \{z \mapsto \ell(h,z) \colon h \in H\}$ *and* $\Delta = \frac{1}{T}\sum_{t=1}^{T-s} d(t+s, T+s)$.

Proof. First, we write $\mathbb{E}[\Phi(\mathbf{Z}_1^T)] \leq \mathbb{E}\left[\sup_{h\in H}(\mathcal{L}_{T+s}(h) - \frac{1}{T}\sum_{t=s}^{T}\ell(h, Z_t))\right] + M\frac{s-1}{T}$. Using the sub-additivity of the supremum, we bound the first term by

$$\mathbb{E}\left[\sup_{h\in H}\frac{1}{T}\sum_{t=1}^{T-s}(\mathcal{L}_{t+s}(h) - \ell(h, Z_{t+s}))\right] + \mathbb{E}\left[\sup_{h\in H}\frac{1}{T}\sum_{t=1}^{T-s}(\mathcal{L}_{T+s}(h) - \mathcal{L}_{t+s}(h))\right].$$

The first summand above is bounded by $2\mathfrak{R}^{\text{seq}}_{T-s}(H_\ell)$ by Theorem 2 of [16]. Note that the result of [16] is for $s = 1$ but it can be extended to an arbitrary s. The second summand is bounded by $\mathbb{E}[\Delta]$ by the definition of the discrepancy. □

McDiarmid's inequality (Corollary 6.10 in [9]), Lemma 7 and Lemma 8 combined yield the following generalization bound for path-dependent error $\mathcal{L}_{T+s}(h)$.

Theorem 2. *Let* L *be a loss function bounded by* M *and let* H *be an arbitrary hypothesis set. Let* $\mathbf{d} = (d_1, \ldots, d_T)$ *with* $d_t = \frac{j_t+1}{T} + \gamma_t\varphi(j_t+2) + \varphi(s)$ *where* $0 \leq j_t \leq T-t$ *and* $\gamma_t = 1$ *iff* $j_t + t < T$ *and* 0 *otherwise (in case training and testing sets are independent we can take* $d_t = \frac{j_t+1}{T} + \gamma_t\varphi(j_t+2)$*). Then, for any* $\delta > 0$*, with probability at least* $1-\delta$*, the following holds for all* $h \in H$*:*

$$\mathcal{L}_{T+s}(h) \leq \frac{1}{T}\sum_{t=1}^{T}\ell(h, Z_t) + \mathbb{E}[\Delta] + 2\mathfrak{R}^{seq}_{T-s}(H_\ell) + M\|\mathbf{d}\|_2\sqrt{2\log\frac{1}{\delta}} + M\frac{s-1}{T}.$$

Observe that for the bound of Theorem 2 to be nontrivial the mixing rate is required to be sufficiently fast. For instance, if $\varphi(\log(T)) = O(T^2)$, then taking $s = \log(T)$ and $j_t = \min\{t, \log T\}$ yields $\|\mathbf{d}\|_2 = O(\sqrt{(\log T)^3/T})$. Combining this with an observation that by Lemma 6, $\mathbb{E}[\Delta] \leq 2\varphi(s) + \frac{1}{T}\sum_{t=1}^{T}\bar{d}(t, T+s)$ one can show that for any $\delta > 0$ with probability at least $1-\delta$, the following holds for all $h \in H$:

$$\mathcal{L}_{T+s}(h) \leq \frac{1}{T}\sum_{t=1}^{T}\ell(h, Z_t) + 2\mathfrak{R}^{\text{seq}}_{T-s}(H_\ell) + \frac{1}{T}\sum_{t=1}^{T}\bar{d}(t, T+s) + O\left(\sqrt{\frac{(\log T)^3}{T}}\right).$$

As commented in Section 3, in general, our bounds are convergent under some natural assumptions examined in the next sections.

5 Convergent Processes

In Section 3 and Section 4 we observed that, for a general non-stationary process, our learning bounds may not converge to zero as a function of the sample size, due to the discrepancies between the training and target distributions. The bounds that we derive suggest that for that convergence to take place, training distributions should "get closer" to the target distribution. However, the issue is that as the sample size grows, the target "is moving". In light of this, we consider a stochastic process that converges to some stationary distribution Π. More precisely, we define

$$\beta(a) = \sup_t \mathbb{E}\big[\|\mathbf{P}_{t+a}(\cdot|\mathbf{Z}_{-\infty}^t) - \Pi\|_{TV}\big] \tag{10}$$

and define $\phi(a)$ in a similar way. We say that a process is β- or ϕ-mixing if $\beta(a) \to 0$ or $\phi(a) \to 0$ as $a \to \infty$ respectively. We remark that this is precisely the mixing assumption used by Agarwal and Duchi [1]. Note that the notions of β- and ϕ-mixing are strictly stronger than the necessary mixing assumptions in Section 3 and Section 4. Indeed, consider a sequence Z_t of independent Gaussian random variables with mean t and unit variance. It is immediate that this sequence is β-mixing but it is not β-mixing. On the other hand, if we use finite-dimensional mixing coefficients, then the following holds:

$$\begin{aligned}
\beta(a) &= \sup_t \mathbb{E}\big[\|\mathbf{P}_{t+a}(\cdot|\mathbf{Z}_{-\infty}^t) - \mathbf{P}_{t+a}\|_{TV}\big] \\
&\leq \sup_t \mathbb{E}\big[\|\mathbf{P}_{t+a}(\cdot|\mathbf{Z}_{-\infty}^t) - \Pi\|_{TV}\big] + \sup_t \sup_A |\mathbb{E}[\mathbb{E}_{t+a}[\mathbf{1}_A|\mathbf{Z}_{-\infty}^t]] - \Pi| \\
&\leq 2\beta(a).
\end{aligned}$$

However, note that a stationary β-mixing process is necessarily β-mixing with $\Pi = \mathbf{P}_0$. We define the *long-term* loss or error $\mathcal{L}_\Pi(h) = \mathbb{E}_\Pi[\ell(h, Z)]$ and observe that $\bar{\mathcal{L}}_T(h) \leq \mathcal{L}_\Pi(h) + M\beta(T)$ since by Lemma 1 the following inequality holds:

$$\begin{aligned}
|\bar{\mathcal{L}}_T(h) - \mathcal{L}_\Pi(h)| &\leq M\|\mathbf{P}_T - \Pi\|_{TV} \leq M\mathbb{E}\big[\|\mathbf{P}_T(\cdot|\mathcal{F}_0) - \Pi\|_{TV}\big] \\
&\leq \sup_t \mathbb{E}\big[\|\mathbf{P}_{T+t}(\cdot|\mathcal{F}_t) - \Pi\|_{TV}\big] = M\beta(T).
\end{aligned}$$

Similarly, we can show that the following holds: $\mathcal{L}_{T+s}(h) \leq \mathcal{L}_\Pi(h) + M\phi(s)$. Therefore, we can use $\mathcal{L}_\Pi$ as a proxy to derive our generalization bound. With this in mind, we consider $\Phi(\mathbf{Z}_1^T)$ defined as in (7) except $\bar{\mathcal{L}}_{T+s}$ is replaced by $\mathcal{L}_\Pi$. Using the sub-sample selection technique of Proposition 2 and the same arguments as in the proof of Lemma 3, we obtain the following result.

Lemma 9. *Let L be a loss function bounded by M and H any hypothesis set. Suppose that $T = ma$ for some $m, a > 0$. Then, for any $\epsilon > \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}_\Pi)]$, the following holds:*

$$\mathbb{P}(\Phi(\mathbf{Z}_1^T) > \epsilon) \leq a\mathbb{P}(\Phi(\widetilde{\mathbf{Z}}_\Pi) - \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}_\Pi)] > \epsilon') + a(m-1)\beta(a), \tag{11}$$

where $\epsilon' = \epsilon - \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}_\Pi)]$ and $\widetilde{\mathbf{Z}}_\Pi$ is an i.i.d. sample of size m from Π.

Using a Rademacher complexity bound [7] for $\mathbb{P}(\Phi(\widetilde{\mathbf{Z}}_{\Pi}) - \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}_{\Pi})] > \epsilon')$ yields the following result.

Theorem 3. *With the assumptions of Lemma 9, for any $\delta > a(m-1)\beta(a)$, with probability $1-\delta$, the following holds for all hypothesis $h \in H$:*

$$\mathcal{L}_{\Pi}(h) \leq \frac{1}{T}\sum_{t=1}^{T} \ell(h, Z_t) + 2\mathfrak{R}_m(H, \Pi) + M\sqrt{\frac{\log \frac{a}{\delta'}}{2m}},$$

where $\delta' = \delta - a(m-1)\beta(a)$ and $\mathfrak{R}_m(H,\Pi) = \frac{1}{m}\mathbb{E}[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i \ell(h, \widetilde{Z}_{\Pi,i})]$ with σ_i a sequence of Rademacher random variables.

Note that our bound requires the confidence parameter δ to be at least $a(m-1)\beta(a)$. Therefore, for the bound to hold with high probability, we need to require $T\beta(a) \to 0$ as $T \to \infty$. This imposes restrictions on the speed of decay of β. Suppose first that our process is algebraically β-mixing, that is $\beta(a) \leq Ca^{-d}$ where $C > 0$ and $d > 0$. Then $T\beta(a) \leq C_0 T a^{-d}$ for some $C_0 > 0$. Therefore, we would require $a = T^{\alpha}$ with $\frac{1}{d} < \alpha \leq 1$, which leads to a convergence rate of the order $\sqrt{T^{(\alpha-1)}\log T}$. Note that we must have $d > 1$. If the processes is exponentially β-mixing, i.e. $\beta(a) \leq Ce^{-da}$ for some $C, d > 0$, then setting $a = \log T^{2/d}$ leads to a convergence rate of the order $\sqrt{T^{-1}(\log T)^2}$.

Finally, we remark that, using the same arguments, it is possible to replace $\mathfrak{R}_m(H,\Pi)$ by its empirical counterpart $\frac{1}{m}\mathbb{E}[\sup_{h\in H}\sum_{t=1}^{T}\sigma_t \ell(h, Z_t)|\mathbf{Z}_1^T]$ leading to data-dependent bounds.

6 Fast Rates for Non-i.i.d. Data

For stationary mixing[8] processes, Steinwart and Christmann [18] have established fast convergence rates when a class of regularized learning algorithms is considered. Agarwal and Duchi [1] also show that stable on-line learning algorithms enjoy faster convergence rates if the loss function is strictly convex. In this section, we present an extension of the local Rademacher complexity results of [2] that imply that under some mild assumptions on the hypothesis set (that are typically used in i.i.d. setting as well) it is possible to have fast learning rates when the data is generated by a convergent process.

The technical assumption that we will exploit is that the Rademacher complexity $\mathfrak{R}_m(H_\ell)$ of the function class $H_\ell = \{z \mapsto \ell(h,z) \colon h \in H\}$ is bounded by some sub-root function $\psi(r)$. A non-negative non-decreasing function $\psi(r)$ is said to be sub-root if $\psi(r)/\sqrt{r}$ is non-increasing. Note that in this section $\mathfrak{R}_m(F)$ always denotes the standard Rademacher complexity with respect to distribution Π defined by $\mathfrak{R}_m(F) = \mathbb{E}[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(\widetilde{Z}_i)]$ where $\widetilde{Z}_i$ is an i.i.d. sample of size m drawn according to Π. Observe that one can always find

[8] In fact, the results of Steinwart and Christmann hold for α-mixing processes which is a weaker statistical assumption then β-mixing.

a sub-root upper bound on $\mathfrak{R}_m(\{f \in F\colon \mathbb{E}[f^2] \leq r\})$ by considering a slightly enlarged function class. More precisely,

$$\mathfrak{R}_m(\{f \in F\colon \mathbb{E}[f^2] \leq r\}) \leq \mathfrak{R}_m(\{g\colon \mathbb{E}[g^2] \leq r, g = \alpha f, \alpha \in [0,1], f \in F\}) = \psi(r)$$

and $\psi(r)$ can be shown to be sub-root (see Lemma 3.4 in [2]). The following analogue of Theorem 3.3 in [2] for the i.i.d. setting is the main result of this section.

Theorem 4. *Let $T = am$ for some $a, m > 0$. Assume that the Rademacher complexity $\mathfrak{R}_m(\{g \in H_\ell\colon \mathbb{E}[g^2] \leq r\})$ is upper bounded by a sub-root function $\psi(r)$ with a fixed point r^*.*[9] *Then, for any $K > 1$ and any $\delta > a(m-1)\beta(a)$, with probability at least $1-\delta$, the following holds for all $h \in H$:*

$$\mathcal{L}_{\Pi}(h) \leq \left(\frac{K}{K-1}\right)\frac{1}{T}\sum_{t=1}^{T}\ell(h, Z_t) + C_1 r^* + \frac{C_2 \log \frac{a}{\delta'}}{m} \tag{12}$$

where $\delta' = \delta - a(m-1)\beta(a)$, $C_1 = 704K/M$, and $C_2 = 26MK + 11M$.

Before we prove Theorem 4, we discuss the consequences of this result. Theorem 4 tells us that with high probability, for any $h \in H$, $\mathcal{L}_{\Pi}(h)$ is bounded by a term proportional to the empirical loss, another term proportional to r^*, which represents the complexity of H, and a term in $O(\frac{1}{m}) = O(\frac{2a}{T})$. Here, m can be thought of as an "effective" size of the sample and a the price to pay for the dependency in the training sample. In certain situations of interest, the complexity term r^* decays at a fast rate. For example, if H_ℓ is a class of $\{0,1\}$-valued functions with finite VC-dimension d, then we can replace r^* in the statement of the Theorem with a term of order $d \log \frac{m}{d}/m$ at the price of slightly worse constants (see Corollary 2.2, Corollary 3.7, and Theorem B.7 in [2]).

Note that unlike standard high probability results, our bound requires the confidence parameter δ to be at least $a(m-1)\beta(a)$. Therefore, for our bound to hold with high probability, we need to require $T\beta(a) \to 0$ as $T \to \infty$ which depends on mixing rate. Suppose that our process is algebraically mixing, that is $\beta(a) \leq Ca^{-d}$ where $C > 0$ and $d > 0$. Then, we can write $T\beta(a) \leq CTa^{-d}$ and in order to guarantee that $T\beta(a) \to 0$ we would require $a = T^\alpha$ with $\frac{1}{d} < \alpha \leq 1$. On the other hand, this leads to a rate of convergence of the order $T^{\alpha-1}\log T$ and in order to have a fast rate, we need $\frac{1}{2} > \alpha$ which is possible only if $d > 2$. We conclude that for a high probability fast rate result, in addition to the technical assumptions on the function class H_ℓ, we may also need to require that the process generating the data be algebraically mixing with exponent $d > 2$. We remark that if the underlying stochastic process is geometrically mixing, that is $\beta(a) \leq Ce^{-da}$ for some $C, d > 0$, then a similar analysis shows that taking $a = \log T^{2/d}$ leads to a high probability fast rate of $T^{-1}(\log T)^2$.

We now present the proof of Theorem 4.

[9] The existence of a unique fixed point is guaranteed by Lemma 3.2 in [2].

Proof. First, we define $\Phi(\mathbf{Z}_1^T) = \sup_{h \in H} \left(\mathcal{L}_{\Pi}(h) - \frac{K}{K-1} \frac{1}{T} \sum_{t=1}^{T} \ell(h, Z_t) \right)$. Using the sub-sample selection technique of Proposition 2, we obtain that $\mathbb{P}(\Phi(\mathbf{Z}_1^T) > \epsilon) \leq a\mathbb{P}(\Phi(\widetilde{\mathbf{Z}}_{\Pi}) > \epsilon) + a(m-1)\beta(a)$, where $\widetilde{\mathbf{Z}}_{\Pi}$ is an i.i.d. sample of size m from Π. By Theorem 3.3 of [2], if $\epsilon = C_1 r^* + \frac{C_2 \log \frac{a}{\delta'}}{m}$, then $a\mathbb{P}(\Phi(\widetilde{\mathbf{Z}}_{\Pi}) > \epsilon)$ is bounded above by $\delta - a(m-1)\beta(a)$, which completes the proof. Note that Theorem 3.3 requires that there exists B such that $\mathbb{E}_{\Pi}[g^2] \leq B\mathbb{E}_{\Pi}[g]$ for all $g \in H_\ell$. This condition is satisfied with $B = M$ since each $g \in H_\ell$ is a bounded non-negative function. □

We remark that, using similar arguments, most of the results of [2] can be extended to the setting of convergent processes. Of course, these results also hold for stationary β-mixing processes since, as we pointed out in Section 5, these are just a special case of convergent processes. However, we note that a slightly tighter bound can be derived for stationary β-mixing processes by using the independent block technique directly instead of relying on the sub-sample selection method.

7 Conclusion

We presented a series of generalization guarantees for learning in presence of non-stationary stochastic processes in terms of an average discrepancy measure that appears as a natural quantity in our general analysis. Our bounds can guide the design of time series prediction algorithms that would tame non-stationarity in the data by minimizing an upper bound on the discrepancy that can be computed from the data [8, 6]. The learning guarantees that we present strictly generalize previous Rademacher complexity guarantees derived for stationary stochastic processes or a drifting setting. We also presented simpler bounds under the natural assumption of convergent processes. In doing so, we have introduced a new sub-sample selection technique that can be of independent interest. Finally, we proved new fast rate learning guarantees in the non-i.i.d. setting. The fast rate guarantees presented can be further expanded by extending in a similar way several of the results of [2].

Acknowledgments

We thank Marius Kloft and Andres Muñoz Medina for discussions about topics related to this research. This work was partly funded by the NSF award IIS-1117591 and the NSERC PGS D3 award.

References

1. Agarwal, A., Duchi, J.C.: The Generalization Ability of Online Algorithms for Dependent Data. IEEE Transactions on Information Theory. 59(1), 573–587 (2013).

2. Bartlett, P.L., Bousquet, O., Mendelson, S.: Local Rademacher complexities. The Annals of Statistics. 33, 1497–1537 (2005).
3. Berti, P., Rigo, P.: A Glivenko-Cantelli theorem for exchangeable random variables. Statistics and Probability Letters. 32, 385–391 (1997).
4. Doukhan, P.: Mixing: Properties and Examples. Lecture Notes in Statistics, vol. 85. Springer Verlag, New York (1989).
5. Eberlein, E.: Weak convergence of partial sums of absolutely regular sequences. Statistics & Probability Letters. 2, 291–293 (1994).
6. Kifer, D., Ben-David, S., & Gehrke, J.: Detecting change in data streams. In: Proceedings of the 30th International Conference on Very Large Data Bases (2004).
7. Koltchinskii, V., Panchenko, D.: Rademacher processes and bounding the risk of function learning. In High Dimensional Probability II, pp. 443-459, Birkhauser (1999).
8. Mansour, Y., Mohri, M., Rostamizadeh, A.: Domain adaptation: learning bounds and algorithms. In: Proceedings of the Annual Conference on Learning Theory (COLT 2009). Omnipress (2009).
9. McDiarmid, C.: On the method of bounded differences. Surveys in Combinatorics, pp. 148–188. Cambridge University Press (1989).
10. Meir, R.: Nonparametric time series prediction through adaptive model selection. Machine Learning. 39(1), 5–34 (2000).
11. Mohri, M., Rostamizadeh, A.: Rademacher complexity bounds for non-i.i.d. processes. In: Advances in Neural Information Processing Systems (NIPS 2008), pp. 1097–1104. MIT Press (2009).
12. Mohri, M., Rostamizadeh, A.: Stability bounds for stationary φ-mixing and β-mixing processes. Journal of Machine Learning. 11 (2010).
13. Mohri, M., Muñoz Medina, A.: New analysis and algorithm for learning with drifting distributions. In: Bshouty, N., Stoltz, G., Vayatis, N., Zeugmann, T. (eds.) ALT 2012. LNCS, vol. 7568, pp 124–138. Springer, Heidelberg (2012).
14. Pestov, V.: Predictive PAC learnability: A paradigm for learning from exchangeable input data. In: 2010 IEEE International Conference on Granular Computing (GrC 2010), pp. 387391, Los Alamitos, California (2010).
15. Rakhlin, A., Sridharan, K., Tewari, A.: Online learning: random averages, combinatorial parameters, and learnability. In: Advances in Neural Information Processing Systems (NIPS 2010), pp. 1984-1992.
16. Rakhlin, A., Sridharan, K., Tewari, A.: Sequential complexities and uniform martingale laws of large numbers. Probability Theory and Related Fields. 1–43 (2014).
17. Shalizi, C.R., Kontorovich. A.: Predictive PAC Learning and Process Decompositions. In: Advances in Neural Information Processing Systems (NIPS 2013), pp. 1619-1627.
18. Steinwart, I., Christmann, A.: Fast learning from non-i.i.d. observations. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C.K.I., Culotta, A. (eds.) Advances in Neural Information Processing Systems (NIPS 2009), pp. 1768–1776. MIT Press (2009).
19. Volkonskii, V.A., Rozanov, Yu A.: Some limit theorems for random functions I. Theory of probability and its applications. 4, 178–197 (1959).
20. Yu, B.: Rates of convergence for empirical processes of stationary mixing sequences. Annals Probability. 22(1), 94–116 (1994).