Machine Learning manuscript No.
(will be inserted by the editor)

Generalization Bounds for Non-stationary Mixing Processes

Vitaly Kuznetsov · Mehryar Mohri

Received: April 4, 2015 / Accepted: February 22, 2016.

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 asymptotically stationary stochastic processes. We also prove that fast learning rates can be achieved by extending existing local Rademacher complexity analyses to the non-i.i.d. setting. We conclude the paper by providing generalization bounds for learning with unbounded losses and non-i.i.d. data.

Keywords Generalization bounds · time series · mixing · non-stationary processes · Markov processes · asymptotic stationarity · fast rates · local Rademacher complexity · unbounded loss

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

An extended abstract of this work appeared as (Kuznetsov and Mohri, 2014).

V. Kuznetsov
Courant Institute of Mathematical Sciences
251 Mercer street, New York, NY 10012, USA
E-mail: vitaly@cims.nyu.edu

M. Mohri
Courant Institute and Google Research
251 Mercer street, New York, NY 10012, USA
E-mail: mohri@cims.nyu.edu

$h\colon \mathcal{X} \to \mathcal{Y}$ 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 (Alquier and Wintenberger, 2010; Alquier et al, 2014; Agarwal and Duchi, 2013; Berti and Rigo, 1997; Shalizi and Kontorovich, 2013; Meir, 2000; Mohri and Rostamizadeh, 2009, 2010; Pestov, 2010; Ralaivola et al, 2010; Steinwart and Christmann, 2009; Yu, 1994). 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 Doukhan (1994), we define β-mixing coefficients for $\mathbf{P}$ as follows. For each positive integer a, we set

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

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 $\|\cdot\|_{\mathrm{TV}}$ between two probability measures P and Q defined on the same σ-algebra of events $\mathcal{G}$ is given by $\|P-Q\|_{\mathrm{TV}} = \sup_{A\in\mathcal{G}} |P(A)-Q(A)|$. 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 dependence with respect to the past weakens over time. We remark that β-mixing coefficients can be defined equivalently as follows:

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

where $\mathbf{P}(\cdot|\cdot)$ denotes conditional probability measure (Doukhan, 1994). Another standard measure of the dependence of the future on the past is the φ-mixing coefficient defined for all $a > 0$ by

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

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, by definition, $\beta(a) \leq \varphi(a)$, so any φ-mixing distribution is necessarily β-mixing. All our results hold for a slightly weaker notion of mixing based on finite-dimensional distributions with $\beta(a) = \sup_t \mathbb{E}\,\|\mathbf{P}_{t+a}(\cdot|\mathbf{Z}_{-\infty}^t) - \mathbf{P}_{t+a}\|_{\mathrm{TV}}$ and $\varphi(a) = \sup_t \sup_{B\in\mathcal{F}_t} \|\mathbf{P}_{t+a}(\cdot|B) - \mathbf{P}_{t+a}\|_{\mathrm{TV}}$. We note that, in certain special cases, such as Markov chains, mixing coefficients admit upper bounds that can be estimated from data (Hsu et al, 2015).

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 $\mathbf{Z}_{t+k}^{t+m+k}$ admit 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 commonly used measure 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

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

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 use the past q 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 $Z_{t+s} = (\mathbf{Y}_{t-q+1}^t, Y_{t+s})$.[1] The generalization ability of stable algorithms with error defined by (4) was studied by Mohri and Rostamizadeh (2010).

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

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

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 (1994) led to VC-dimension bounds for $\bar{\mathcal{L}}_{T+s}$ under the assumption of stationarity and β-mixing. Later, Meir (2000) used that to derive generalization bounds in terms of covering numbers of H. These results have been further extended by Mohri and Rostamizadeh (2009) to data-dependent learning bounds in terms of the Rademacher complexity of H. Ralaivola et al (2010), Alquier and Wintenberger (2010), and Alquier et al (2014) provide PAC-Bayesian learning bounds under the same assumptions.

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 (2013), who present bounds for stable on-line learning algorithms under the assumptions of asymptotically stationary process.[2] 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.[3] We also show that mixing is in fact necessary for learning with $\bar{\mathcal{L}}_{T+s}$, which further motivates the study of $\mathcal{L}_{T+s}$.

Next, we strengthen our assumptions and give generalization bounds for asymptotically stationary processes. In doing so, we provide guarantees for learning with Markov chains - most widely used class of stochastic processes. These results are

[1] 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 (2000)) to derive generalization bounds for unbounded-memory models.

[2] Agarwal and Duchi (2013) additionally assume that distributions are absolutely continuous and that the loss function is convex and Lipschitz.

[3] While this work was under review, Kuznetsov and Mohri (2015) used techniques that appeared in the extended abstract of this work (Kuznetsov and Mohri, 2014) to establish generalization bounds for learning with non-stationary non-mixing processes.

algorithm-agnostic analogues of the algorithm-dependent bounds of Agarwal and Duchi (2013). Agarwal and Duchi (2013) also prove fast convergence rates when a strongly convex loss is used. Similarly, Steinwart and Christmann (2009) showed that regularized learning algorithms admit faster convergence rates under the assumptions of mixing and stationarity. We show that this is in fact a general phenomenon and use local Rademacher complexity techniques (Bartlett et al, 2005) to establish faster convergence rates for stationary mixing or asymptotically stationary processes.

Finally, all the existing learning guarantees only hold for bounded loss functions. However, for a large class of time series prediction problems, this assumption is not valid. We conclude this paper by providing the first learning guarantees for unbounded losses and non-i.i.d. data.

A key ingredient of the bounds we present is the notion of *discrepancy* between two probability distributions that was used by Mohri and Muñoz (2012) to give generalization bounds for sequences of independent (but not identically distributed) random variables. In our setting, discrepancy can be defined as

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

Similarly, we define $\bar{d}(t_1, t_2)$ by replacing $\mathcal{L}_t$ with $\bar{\mathcal{L}}_t$ in the definition of $d(t_1, t_2)$. Discrepancy is a natural measure of the non-stationarity of a stochastic process with respect to 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. 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 $t_1 - t_2$. If L is a squared loss and a set of linear hypothesis $H = \{\mathbf{Y}_{t-q+1}^T \mapsto w \cdot \mathbf{Y}_{t-q+1}^T \colon w \in \mathbb{R}^q\}$ is used, then it can be shown (see Lemma 12 Appendix A) that in this case we again have $\bar{d}(t_1, t_2) = 0$ for all $t_1, t_2 \in \mathbb{Z}$. This example highlights the fact that discrepancy captures not only properties of the distribution of the stochastic processes, but also properties of other important components of the learning problem such as the hypothesis set H and the loss function L. 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 (Mansour et al, 2009; Kifer et al, 2004).

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 5 present learning guarantees for averaged and path-dependent errors respectively. In Section 4 we establish that mixing is a necessary condition for learning with averaged path-dependent errors. In Section 6, we analyze generalization with asymptotically stationary processes. We present fast learning rates for the non-i.i.d. setting in Section 7. In Section 8, we conclude with generalization bounds for unbounded loss functions.

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 known concentration results. One way

to achieve this is via the independent block technique introduced by Bernstein (1927) 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 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 (Yu, 1994), which in turn is based on (Eberlein, 1984) and (Volkonskii and Rozanov, 1959).[4] For the sake of completeness, we present the full proof of this result below. We will also use the main steps of this proof as stand-alone results later in the sequel.

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*

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

Proof We start by proving this claim for simple functions of the form

$$h = \sum_{j=1}^{k} c_j \mathbf{1}_{A_j}, \tag{7}$$

where A_js are in $\mathcal{F}$ and pairwise disjoint. Note that we do not require $c_j \geq 0$. Observe that in this case

$$\begin{aligned}\mathop{\mathbb{E}}_{Q} h - \mathop{\mathbb{E}}_{P} h &= \sum_{j=1}^{k} c_j (Q(A_j) - P(A_j)) \\ &\leq \sum_{j \in J_1} c_j (Q(A_j) - P(A_j)) + \sum_{j \in J_2} c_j (Q(A_j) - P(A_j))\end{aligned}$$

[4] The bound stated in (Yu, 1994) only holds in case $M_1 = 0$, i.e. for non-negative g and $a_t = a$ for all t. Indeed, to see that if $|g| \leq M$ it need not be the case that $|\,\mathbb{E}[g(\widetilde{\mathbf{Z}}^o)] - \mathbb{E}[g(\mathbf{Z}^o)]| \leq M(m-1)\beta(a)$, consider $Z_t = Z$ for all t, where $\mathbb{P}(Z = 1) = p$ and $\mathbb{P}(Z = -1) = q$. Suppose $g\colon \mathbb{R}^T \to \mathbb{R}$ s.t. $g(z_1, \ldots, z_{ma}) = 1$ if $z_1 = \ldots = z_{ma}$ and -1 otherwise. Then one can show that $\mathbb{E}\, g(S_1) - \mathbb{E}\, g(\widetilde{S}_1) = 2 - 2(p^m + (1-p)^m$ and $\beta(a) = p(1-p)$ for any a. For any m we can find p such that $2 - 2(p^m + (1-p)^m > (m-1)p(1-p)$. For instance, if $m = 2$ then $p = \frac{1}{2}$ will suffice.

where $J_1 = \{j \colon (Q(A_j) - P(A_j)) \leq 0, c_j \leq 0\}$ and $J_2 = \{j \colon (Q(A_j) - P(A_j)) \geq 0, c_j \geq 0\}$. Therefore,

$$\begin{aligned}\mathop{\mathbb{E}}_{Q} h - \mathop{\mathbb{E}}_{P} h &\leq M_1 \sum_{j \in J_1} (P(A_j) - Q(A_j)) + M_2 \sum_{j \in J_2} (Q(A_j) - P(A_j)) \\ &= M_1 \Big(P(\cup_{j \in J_1} A_j) - Q(\cup_{j \in J_1} A_j) \Big) + M_2 \Big(Q(\cup_{j \in J_2} A_j) - P(\cup_{j \in J_2} A_j) \Big) \\ &\leq (M_1 + M_2) \|Q - P\|_{\mathrm{TV}},\end{aligned}$$

where the equality follows from the fact that A_js are disjoint. By symmetry, $\mathbb{E}_P h - \mathbb{E}_Q h \leq (M_1 + M_2)\|Q - P\|_{\mathrm{TV}}$ and combining these results shows that the lemma holds for all simple functions of the form (7). To complete the proof of the lemma we use a standard approximation argument. Set $\Psi_n(x) = \min(n, 2^{-n} \lfloor 2^n x \rfloor)$ for $x \geq 0$ and $\Psi_n(x) = -\min(n, 2^{-n} \lfloor -2^n x \rfloor)$ for $x < 0$. From this definition it is immediate that $\Psi_n(h)$ converges pointwise to h as $n \to \infty$ and $-M_1 \leq \Psi_n(h) \leq M_2$. Therefore, by the bounded convergence theorem, for any $\epsilon > 0$, we can find n such that $|\mathbb{E}_P h - \mathbb{E}_P \Psi_n(h)| < \epsilon$ and $|\mathbb{E}_Q h - \mathbb{E}_Q \Psi_n(h)| < \epsilon$. Since $\Psi_n(h)$ is a simple function of the form (7), by our previous result and the triangle inequality, we find that

$$\begin{aligned}|\mathop{\mathbb{E}}_{P} h - \mathop{\mathbb{E}}_{Q} h| &\leq |\mathop{\mathbb{E}}_{P} h - \mathop{\mathbb{E}}_{P} \Psi_n(h)| + |\mathop{\mathbb{E}}_{P} \Psi_n(h) - \mathop{\mathbb{E}}_{Q} \Psi_n(h)| + |\mathop{\mathbb{E}}_{Q} h - \mathop{\mathbb{E}}_{Q} \Psi_n(h)| \\ &\leq 2\epsilon + (M_1 + M_2)\|Q - P\|_{\mathrm{TV}}.\end{aligned}$$

Since the inequality holds for all $\epsilon > 0$, we conclude that $|\mathbb{E}_P h - \mathbb{E}_Q h| \leq (M_1 + M_2)\|Q - P\|_{\mathrm{TV}}$. □

Note that, if $|g| < M$, then $\|\mathbb{E}_Q g - \mathbb{E}_P g\| \leq 2M\|P - Q\|_{\mathrm{TV}}$ and the factor of 2 is necessary in this bound. Consider a measure space $\Omega = \{0, 1\}$ equipped with a σ-algebra $\mathcal{F} = \{\emptyset, \{0\}, \{1\}, \Omega\}$. Let Q and P be probability measures on $(\Omega, \mathcal{F})$ such that $Q\{0\} = P\{1\} = 1$ and $Q\{1\} = P\{0\} = 0$. If $h(0) = 1$ and $h(1) = -1$ then $|\mathbb{E}_Q h - \mathbb{E}_P h| = 2 > 1 = \|P - Q\|_{\mathrm{TV}}$. Lemma 1 extended via induction yields the following result.

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*

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

for $j \geq 1$ and $\beta_0 = \|P_1 - Q_1\|_{\mathrm{TV}}$. 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

$$|\mathop{\mathbb{E}}_{P}[h] - \mathop{\mathbb{E}}_{Q}[h]| \leq (M_1 + M_2) \sum_{j=0}^{m-1} \beta_j$$

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

Proof We will prove this claim by induction on m. First suppose $m = 1$. Then, the conclusion follows from Lemma 1. Next, assume that the claim holds for $m-1$, where $m \geq 2$. We will show that it must also hold for m. Observe that

$$|\mathop{\mathbb{E}}_{P} h - \mathop{\mathbb{E}}_{Q} h| \leq |\mathop{\mathbb{E}}_{P} h - \mathop{\mathbb{E}}_{P_{m-1}\otimes Q_m} h| + |\mathop{\mathbb{E}}_{P_{m-1}\otimes Q_m} h - \mathop{\mathbb{E}}_{Q_1\otimes\ldots\otimes Q_m} h|.$$

For the first term we observe that

$$\begin{aligned}|\mathop{\mathbb{E}}_{P} h - \mathop{\mathbb{E}}_{P_{m-1}\otimes Q_m} h| &= |\mathop{\mathbb{E}}_{P_{m-1}}\mathop{\mathbb{E}}_{P_m(\cdot|\mathcal{G}_{m-1})} h - \mathop{\mathbb{E}}_{P_{m-1}}\mathop{\mathbb{E}}_{Q_m} h| \\ &\leq \mathop{\mathbb{E}}_{P_{m-1}} |\mathop{\mathbb{E}}_{P_m(\cdot|\mathcal{G}_{m-1})} h - \mathop{\mathbb{E}}_{Q_m} h|,\end{aligned}$$

where $\mathcal{G}_j = \prod_{k=1}^{j} \mathcal{F}_k$. Applying Lemma 1 we have that the first term is bounded by $(M_1+M_2)\beta_{m-1}$. To bound the second term we apply Fubini's Theorem, Lemma 1 and inductive hypothesis to get that

$$\begin{aligned}|\mathop{\mathbb{E}}_{P_{m-1}\otimes Q_m} h - \mathop{\mathbb{E}}_{Q_1\otimes\ldots\otimes Q_m} h| &= |\mathop{\mathbb{E}}_{Q_m}\mathop{\mathbb{E}}_{P_{m-1}} h - \mathop{\mathbb{E}}_{Q_m}\mathop{\mathbb{E}}_{Q_1\otimes\ldots\otimes Q_{m-1}} h| \\ &\leq \mathop{\mathbb{E}}_{Q_m} |\mathop{\mathbb{E}}_{P_{m-1}} h - \mathop{\mathbb{E}}_{Q_1\otimes\ldots\otimes Q_{m-1}} h| \\ &\leq (M_1 + M_2)\sum_{j=0}^{m-2}\beta_j\end{aligned}$$

and the desired conclusion follows. □

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.

Proof (Proof of Proposition 1) We start by establishing some notation. Let P_j denote the joint distribution of $\mathbf{Z}(1), \mathbf{Z}(3), \ldots, \mathbf{Z}(2j-1)$ and let Q_j denote the distribution of $\mathbf{Z}(2j-1)$ (or equivalently $\widetilde{\mathbf{Z}}(2j-1)$). We will also denote the joint distribution of $\mathbf{Z}(2j+1), \ldots, \mathbf{Z}(2m-1)$ by P^j. Set $P = P_m$ and $Q = Q_1 \otimes \ldots \otimes Q_m$. In other words, P and Q are distributions of $\mathbf{Z}^o$ and $\widetilde{\mathbf{Z}}^o$ respectively. Then

$$|\,\mathbb{E}\, g(\widetilde{\mathbf{Z}}^o) - \mathbb{E}\, g(\mathbf{Z}^o)| = |\mathop{\mathbb{E}}_{Q} g - \mathop{\mathbb{E}}_{P} g| \leq (M_1 + M_2)\sum_{j=0}^{m-1}\beta_j$$

by Lemma 2. Observing that $\beta_j \leq \beta(a_{2j})$ and $\beta_0 = 0$ completes the proof of the Proposition 1. □

Proposition 1 is not the only way to relate mixing and independent cases. Next, we introduce an alternative technique that we name *sub-sample selection*, which is particularly useful when the process is asymptotically stationary. 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*

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

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

The proof of Proposition 2 is the same as the proof of Proposition 1 modulo the definition of measure Q which we set to Π^m. Proposition 2 is commonly applied with Π the stationary probability measure of an asymptotically stationary 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, we define $\Phi(\mathbf{Z}_1^T)$ as follows:

$$\Phi(\mathbf{Z}_1^T) = \sup_{h \in H} \left(\bar{\mathcal{L}}_{T+s}(h) - \frac{1}{T} \sum_{t=1}^{T} \ell(h, Z_t) \right). \tag{8}$$

We assume that Φ is measurable which can be guaranteed under some additional mild assumption on $\mathcal{Z}$ and H. 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:*

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

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:*

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

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

Proof We apply McDiarmid's inequality (McDiarmid, 1989) 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 the *block* Rademacher complexity defined by

$$\mathfrak{R}(\widetilde{\mathbf{Z}}^o) = \frac{1}{|I_1|}\,\mathbb{E}\left[\sup_{h\in H} \sum_{i=1}^{m} \sigma_i\, l\big(h, \mathbf{Z}(2i-1)\big)\right], \tag{9}$$

where σ_i is a sequence of Rademacher random variables and $l(h, \mathbf{Z}(2i-1)) = \sum_t \ell(h, Z_t)$ and where the sum is taken over t in the ith odd block. 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:*

$$\max(\mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^o)], \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}^e)]) \leq 2\max(\mathfrak{R}(\widetilde{\mathbf{Z}}^o), \mathfrak{R}(\widetilde{\mathbf{Z}}^e)) + \max(\Delta^1, \Delta^2). \tag{10}$$

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

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

The second term can be written as

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

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

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

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 $(T/(\|\mathbf{a}^e\|_2 + \|\mathbf{a}^o\|_2))^2$. 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 6 and Section 7, we will describe some natural assumptions under which this convergence does occur. However, in general, a small discrepancy is necessary for learning to be possible, since Barve and Long (1996) showed that $O(\gamma^{1/3})$ is a lower bound on the generalization error in the setting of binary classification where the sequence $\mathbf{Z}_1^T$ is a sequence of independent but not identically distributed random variables and where γ is an upper bound on discrepancy. We also note that Theorem 1 can be stated in terms of a slightly tighter notion of discrepancy $\sup_h |\bar{\mathcal{L}}_{T+s} - (1/|I_j|)\sum_{t\in I_j}\bar{\mathcal{L}}_t|$ instead of average *instantaneous* discrepancies Δ^j.

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 sub-samples $\mathbf{Z}^{(j)}$ 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.

Corollary 1 *With the assumptions of Lemma 3, and $T = 2am$, for some $a, m > 0$, for any $\delta > 2(m-1)\beta(a)$, with probability $1-\delta$, the following holds for all hypotheses $h \in H$:*

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

If the process is stationary, then we recover as a special case the generalization bound of Mohri and Rostamizadeh (2009). If $\mathbf{Z}_1^T$ is a sequence of independent but not identically distributed random variables, we recover the results of Mohri and Muñoz (2012). In the i.i.d. case, Theorem 1 reduces to the generalization bounds of Koltchinskii and Panchenko (2000).

The Rademacher complexity $\mathfrak{R}_m(\widetilde{\mathbf{Z}}^{(j)})$ that appears in our bound is not standard. In particular, the random variables $\widetilde{Z}_j, \widetilde{Z}_{2a+j}, \ldots, \widetilde{Z}_{2a(m-1)+j}$ may follow different distributions. However, $\mathfrak{R}_m(\widetilde{\mathbf{Z}}^{(j)})$ can be analyzed in the same way as the standard Rademacher complexity defined in terms of i.i.d. sample. For instance, it can be bounded in terms of distribution-agnostic combinatorial complexity measures such as the VC-dimension or growth function using standard results such as Massart's lemma (Mohri et al, 2012). Alternatively, for ρ-Lipschitz losses, Talagrand's contraction principle can be used to bound the Rademacher complexity of the set of linear hypotheses $H = \{x \to \mathbf{w} \cdot \Psi(x) \colon \|\mathbf{w}\|_{\mathcal{H}} \leq \Lambda\}$ by $\rho r \Lambda / \sqrt{m}$, where $\mathcal{H}$ is a Hilbert space associated to the feature map Ψ and kernel K and where $r = \sup_x K(x,x)$.

4 Mixing and Averaged Generalization Error

In this section, we show that mixing is in fact necessary for generalization with respect to averaged error.

We consider a task of forecasting binary sequences over $\mathcal{Y} = \{\pm 1\}$ using side information in $\mathcal{X}$ and history of the stochastic process. That is, a learning algorithm $\mathcal{A}$ is provided with a sample $\mathbf{Z}_1^T \in \mathcal{X}^T \times \{\pm 1\}^T$ and produces a hypothesis $h_{\mathbf{Z}_1^T}$. At time $T+1$, side information X_{T+1} is observed and $h_{\mathbf{Z}_1^T}(X_{T+1})$ is forecasted by the algorithm. The performance of the algorithm is evaluated using $L(y, y') = \mathbf{1}_{y \neq y'}$.

We have the following result.

Theorem 2 *Let H be a set of hypotheses with $d = \text{VC-dim}(H) \geq 2$. For any algorithm $\mathcal{A}$, there is a stationary process that is not β-mixing and such that for each T, there is $T' \geq T$ such that*

$$\mathbb{P}\left(\bar{\mathcal{L}}_{T'+1}(h_{\mathbf{Z}_1^{T'}}) - \inf_{h \in H} \bar{\mathcal{L}}_{T'+1}(h) \geq \frac{1}{2}\right) \geq \frac{1}{8}. \tag{11}$$

Proof Since $d \geq 2$, there is $\mathcal{X}' = \{x_1, x_2\} \subset \mathcal{X}$ such that this set if fully shattered, that is each of the dichotomies is possible on this set. The stochastic process we will define admits $\mathcal{X}'$ for support. We will further assume $H = H'$, where $H' = \{h_1, h_2, h_3, h_4\}$ is a set of hypotheses that represent all possible dichotomies on $\mathcal{X}$.

Now let S_T be sample of size T drawn i.i.d. from a Dirac mass $\delta_{(x_1,1)}$ and let h_{S_T} be a hypothesis produced by $\mathcal{A}$ when trained on this sample. Note that h_{S_T} is a random variable and the randomness may come from two sources: the sample S_T and the algorithm $\mathcal{A}$ itself. Thus, conditioned on S_T, let p_T be the distribution over H used by the algorithm to produce h_{S_T}. Note that p_T is completely determined by $(x_1, 1, T)$. If the algorithm is deterministic then p_T is a point mass.

Consider now a sequence of distributions $p_1, p_2, \ldots$, define

$$h_T = \operatorname*{argmax}_{h \in H} p_T(h)$$

and observe that $p_T(h_T) \geq \frac{1}{4}$. Let h^* be an element of H that appears in the sequence $h_1, h_2, \ldots$ infinitely often. The existence of h^* is guaranteed by the finiteness of H.

Let $y_0 = -h^*(x_2)$. We define a distribution $\mathcal{D} = \frac{1}{2}\delta_{(x_1,1)} + \frac{1}{2}\delta_{(x_2,y_0)}$. Then let $(X_1, Y_1) \sim \mathcal{D}$ and for all $t > 1$,

$$(X_t, Y_t) \sim \begin{cases} \delta_{(x_1,1)}, & \text{if } X_1 = x_1, \\ \delta_{(x_2,y_0)}, & \text{otherwise.} \end{cases}$$

We first show that this stochastic process satisfies (11). Indeed, observe that $\inf_{h \in H} \bar{\mathcal{L}}_{T'+1}(h) = 0$ and if $E_T = \{\bar{\mathcal{L}}_{T+1}(h_{\mathbf{Z}_1^T}) \geq \frac{1}{2}\}$

$$\mathbb{P}(E_{T'}) = \frac{1}{2}\mathbb{P}(E_{T'}|X_1 = x_1) + \frac{1}{2}\mathbb{P}(E_{T'}|X_1 \neq x_1) \geq \frac{1}{2}\mathbb{P}(E_{T'}|X_1 = x_1).$$

Choose T' such that $h_{T'} = h^*$ and observe that in that case

$$\frac{1}{2}\mathbb{P}(E_{T'}|X_1 = x_1) \geq \frac{1}{8}\mathbb{P}(E_{T'}|h^* = h_{T'} = h_{\mathbf{Z}_1^T}, X_1 = x_1) = \frac{1}{8},$$

where the last equality follows from:

$$\bar{\mathcal{L}}_{T+1}(h_{\mathbf{Z}_1^T}) = \frac{1}{2}L(h_{T'}(x_1), 1) + \frac{1}{2}L(h_{T'}(x_2), -h_{T'}(x_2)) \geq \frac{1}{2},$$

when we condition on $h^* = h_{T'} = h_{\mathbf{Z}_1^T}$ and $X_1 = x_1$.

We conclude this proof by showing that this process is stationary and not β-mixing. One can check that for any t, and any k and any sequence $(z_1, \ldots, z_k)$, the following holds

$$\mathbb{P}(Z_t = z_1, \ldots, Z_{t+k} = z_k) = \begin{cases} \frac{1}{2}, & \text{if } z_1 = \ldots = z_k = (x_1, 1), \\ \frac{1}{2}, & \text{if } z_1 = \ldots = z_k = (x_2, y_0), \\ 0, & \textit{otherwise.} \end{cases}$$

Since the right-hand side is independent of t it follows that this process is stationary. Now observe that for any event A

$$|\mathbf{P}_{t+a}(A|Z_1 = (x_1, 1), \mathbf{Z}_2^T) - \mathbf{P}_{t+a}(A)| = \frac{1}{2}|\delta_{(x_2,y_0)}(A) - \delta_{(x_1,1)}(A)|$$

and taking the supremum over A yields that $\|\mathbf{P}_{t+a}(\cdot|Z_1 = (x_1, 1), \mathbf{Z}_2^T) - \mathbf{P}_{t+a}\|_{\mathrm{TV}} = \frac{1}{2}$. Similarly, one can show that $\|\mathbf{P}_{t+a}(\cdot|Z_1 = (x_2, y_0), \mathbf{Z}_2^T) - \mathbf{P}_{t+a}\|_{\mathrm{TV}} = \frac{1}{2}$, which proves that $\beta(a) = \frac{1}{2}$ for all a and this process is not β-mixing. □

We note that, in fact, the process that is constructed in Theorem 2 is not even α-mixing.

Note that this result does not imply that mixing is necessary for generalization with respect to path-dependent generalization error and this further motivates the study of this quantity.

5 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. In this section, we will use $\Phi(\mathbf{Z}_1^T)$ to denote the same quantity as in (8) 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 (McDiarmid, 1989) or Appendix C). 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 $g\colon \mathcal{Z}^{k+j} \to \mathbb{R}$ 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:

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

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

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

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 needed 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 (Rakhlin et al, 2010) and further shown to characterize learning in scenarios with sequential data (Rakhlin et al, 2011b,a, 2015). We give a brief overview of sequential Rademacher complexity in Appendix B.

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 (Rakhlin et al, 2015). Note that the result of Rakhlin et al (2015) is for $s=1$ but it can be extended to an arbitrary s. We explain how to carry out this extension in Appendix B. The second summand is bounded by $\mathbb{E}[\Delta]$ by the definition of the discrepancy. □

Note that Lemma 8 and all subsequent results in this Section can be stated in terms of a slightly tighter notion of discrepancy $\mathbb{E}[\sup_h|\mathcal{L}_{T+s}-(1/T)\sum_{t=1}^{T}\mathcal{L}_t|]$ instead of average instantaneous discrepancy $\mathbb{E}[\Delta]$.

McDiarmid's inequality (Corollary 6.10 in (McDiarmid, 1989)), Lemma 7 and Lemma 8 combined yield the following generalization bound for path-dependent error $\mathcal{L}_{T+s}(h)$.

Theorem 3 *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 3 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.

6 Asymptotically Stationary Processes

In Section 3 and Section 5 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 $\mathit{\Pi}$. More precisely, we define

$$\beta(a) = \sup_t \mathbb{E}\left[\|\mathbf{P}_{t+a}(\cdot|\mathbf{Z}^t_{-\infty}) - \mathit{\Pi}\|_{\text{TV}}\right] \tag{12}$$

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 define a process to be *asymptotically stationary* if it is either β- or ϕ-mixing.[5] This is precisely the assumption used by Agarwal and Duchi (2013) to give stability bounds for on-line learning algorithms. Note that the notions of β- and ϕ-mixing are strictly stronger than the necessary mixing assumptions in Section 3 and Section 5. 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:

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

[5] Note that asymptotically stationary processes are called convergent (Kuznetsov and Mohri, 2014).

However, note that a stationary β-mixing process is necessarily β-mixing with $\Pi = \mathbf{P}_0$.

Asymptotically stationary processes constitute an important class of stochastic processes that are common in many modern applications. In particular, any homogeneous aperiodic irreducible Markov chain with stationary distribution Π is asymptotically stationary since

$$\phi(a) = \sup_t \sup_{z_1^T} \left[\|\mathbf{P}_{t+a}(\cdot|z_1^T) - \Pi\|_{\mathrm{TV}}\right] = \sup_{z \in \mathcal{Z}} \left[\|\mathbf{P}_a(\cdot|z) - \Pi\|_{\mathrm{TV}}\right] \to 0,$$

where the second equality follows from homogeneity and the Markov property and where the limit result is a consequence of the Markov Chain Convergence Theorem. Note that, in general, a Markov chain may not be stationary, which shows that the generalization bounds that we present here are an important extension of statistical learning theory to a scenario frequent appearing in applications.

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:

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

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 (8) 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:*

$$\mathbb{P}(\Phi(\mathbf{Z}_1^T) > \epsilon) \leq a\mathbb{P}(\Phi(\widetilde{\mathbf{Z}}_\Pi) - \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}_\Pi)] > \epsilon') + T\beta(a), \tag{13}$$

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.

Proof By convexity of the supremum, the following holds:

$$\Phi(\mathbf{Z}_1^T) \leq \frac{1}{a}\sum_{j=1}^{a} \sup_{h \in H} \Big(\mathcal{L}_\Pi(h) - \frac{1}{m}\sum_{t=0}^{m-1} \ell(h, Z_{ta+j})\Big).$$

We denote by $\Phi(\mathbf{Z}^{(j)})$ the j-summand appearing on the right-hand side. For $\Phi(\mathbf{Z}_1^T)$ to exceed ϵ at least one of $\Phi(\mathbf{Z}^{(j)})$s must exceed ϵ. Thus, by the union bound, we have that

$$\mathbb{P}(\Phi(\mathbf{Z}_1^T) \geq \epsilon) \leq \sum_{j=1}^{a} \mathbb{P}(\Phi(\mathbf{Z}^{(j)}) \geq \epsilon).$$

Applying Proposition 2 to each term on the right-hand side yields the desired result. □

Using the standard Rademacher complexity bound of Koltchinskii and Panchenko (2000) for $\mathbb{P}(\Phi(\widetilde{\mathbf{Z}}_\Pi) - \mathbb{E}[\Phi(\widetilde{\mathbf{Z}}_\Pi)] > \epsilon')$ yields the following result.

Theorem 4 *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 $\delta' = \delta - T\beta(a)$ and $\mathfrak{R}_m(H,\Pi) = \frac{1}{m}\,\mathbb{E}\left[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i \ell(h, \widetilde{Z}_{\Pi,i})\right]$ with σ_i a sequence of Rademacher random variables.

Note that our bound requires the confidence parameter δ to be at least $T\beta(a)$. 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 $\sqrt{T^{-1}(\log T)^2}$.

The Rademacher complexity $\mathfrak{R}_m(H,\Pi)$ can be upper bounded by distribution-agnostic combinatorial measures of complexity such as VC-dimension using standard techniques. Alternatively, using the same arguments, it is possible to replace $\mathfrak{R}_m(H,\Pi)$ by its empirical counterpart $\frac{1}{m}\,\mathbb{E}[\sup_{h\in H}\sum_{t=0}^{m-1}\sigma_t \ell(h, Z_{at+j})|\mathbf{Z}^{(j)}]$ leading to data-dependent bounds.

Corollary 2 *With the assumptions of Lemma 9, for any $\delta > 2a(m-1)\beta(a)$, with probability $1-\delta$, the following holds for all hypothesis $h \in H$:*

$$\mathcal{L}_{\Pi}(h) \leq \frac{1}{T}\sum_{t=1}^{T} \ell(h, Z_t) + \frac{2}{a}\sum_{j=1}^{a} \widehat{\mathfrak{R}}_m(H, \mathbf{Z}^{(j)}) + 3M\sqrt{\frac{\log \frac{2a}{\delta'}}{2m}},$$

where $\delta' = \delta - T\beta(a)$ and $\widehat{\mathfrak{R}}_m(H,\mathbf{Z}^{(j)}) = \frac{1}{m}\,\mathbb{E}\left[\sup_{h\in H}\sum_{t=0}^{m-1}\sigma_t \ell(h, Z_{at+j})|\mathbf{Z}^{(j)}\right]$ is empirical Rademacher complexity with σ_i a sequence of Rademacher random variables.

Proof By union bound, it follows that

$$\mathbb{P}\Big(\mathfrak{R}_m(H,\Pi) - \frac{1}{a}\sum_{j=1}^{a}\widehat{\mathfrak{R}}_m(H,\mathbf{Z}^{(j)}) \geq \epsilon\Big) \leq \sum_{j=1}^{a}\mathbb{P}(\Psi(\mathbf{Z}^{(j)}) \geq \epsilon),$$

where $\Psi(\mathbf{Z}^{(j)}) = \mathfrak{R}_m(H,\Pi) - \widehat{\mathfrak{R}}_m(H,\mathbf{Z}^{(j)})$. By Proposition 2, we can bound the above by

$$a\mathbb{P}(\Psi(\mathbf{Z}_{\Pi}) \geq \epsilon) + T\beta(a),$$

where $\mathbf{Z}_{\Pi}$ is an i.i.d. sample of size m from Π. The rest of the proof follows from the standard result for Rademacher complexity of i.i.d. random variables, McDiarmid's inequality and union bound. □

The significance of Corollary 2 follows from the fact that $\widehat{\mathfrak{R}}_m(H, \mathbf{Z}^{(j)})$ can be estimated from the sample $\mathbf{Z}_1^T$ leading to learning bounds that can computed from the data.

We conclude this section by observing that Theorem 1 and Theorem 3 could also be used to derive similar learning guarantees to the ones presented in this section by directly upper bounding the discrepancy:

$$\begin{aligned}\bar{d}(T+s,t) = \sup_h \left|\bar{\mathcal{L}}_{T+s}(h) - \bar{\mathcal{L}}_t(h)\right| \leq & \sup_h \left|\bar{\mathcal{L}}_{T+s}(h) - \mathcal{L}_{\Pi}(h)\right| + \sup_h \left|\bar{\mathcal{L}}_t(h) - \mathcal{L}_{\Pi}(h)\right| \\ \leq & \mathbb{E}[\sup_h \left| \mathbb{E}[\ell(h, Z_{T+s})|\mathbf{Z}_{-\infty}^0] - \mathcal{L}_{\Pi}(h)\right| \\ & + \mathbb{E}[\sup_h \left| \mathbb{E}[\ell(h, Z_t)|\mathbf{Z}_{-\infty}^0] - \mathcal{L}_{\Pi}(h)\right| \\ \leq & \beta(T+s) + \beta(t),\end{aligned}$$

and similarly for $d(T+s,t) \leq \phi(T+s) + \phi(t) + 2\phi(s)$. However, we chose to illustrate our sub-sample selection technique in this simpler setting since we will use it in Section 7 and Section 8 to give fast rates and learning guarantees for unbounded losses for non-i.i.d. data.

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

For stationary mixing processes, Steinwart and Christmann (2009) established fast convergence rates when a class of regularized learning algorithms is considered.[6] Agarwal and Duchi (2013) also showed 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 Bartlett et al (2005) which imply that, under some mild assumptions on the hypothesis set (that are typically adopted in i.i.d. setting as well), it is possible to achieve fast learning rates when the data is generated by an asymptotically stationary 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 and F is an arbitrary function class. Observe that one can always find 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, for

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

$\psi(r)$ can be shown to be sub-root (see Lemma 3.4 in (Bartlett et al, 2005)). The following analogue of Theorem 3.3 in (Bartlett et al, 2005) for the i.i.d. setting is the main result of this section.

[6] In fact, the results of Steinwart and Christmann (2009) hold for α-mixing processes which is a weaker statistical assumption than β-mixing.

Theorem 5 *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^*.*[7] *Then, for any $K > 1$ and any $\delta > T\beta(a)$, with probability at least $1 - \delta$, the following holds for all $h \in H$:*

$$\mathcal{L}_{\Pi}(h) \leq \left(\frac{K}{K-1}\right)\frac{1}{T}\sum_{t=1}^{T}\ell(h, Z_t) + C_1 r^* + \frac{C_2 \log \frac{a}{\delta'}}{m} \tag{14}$$

where $\delta' = \delta - T\beta(a)$, $C_1 = 704K/M$, and $C_2 = 26MK + 11M$.

Before we prove Theorem 5, we discuss the consequences of this result. Theorem 5 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 (Bartlett et al, 2005)).

Note that unlike standard high probability results, our bound requires the confidence parameter δ to be at least $T\beta(a)$. 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 achieve 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 5.

Proof First, we define

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

Observe that $\Phi(\mathbf{Z}_1^T) \leq \frac{1}{a}\sum_{j=1}^{a}\Phi(\mathbf{Z}^{(j)})$ and at least one of $\Phi(\mathbf{Z}^{(j)})$s must exceed ϵ in order for event $\{\Phi(\mathbf{Z}_1^T) \geq \epsilon\}$ to occur. Therefore, by the union bound and the sub-sample selection technique of Proposition 2, we obtain that

$$\mathbb{P}(\Phi(\mathbf{Z}_1^T) > \epsilon) \leq a\mathbb{P}(\Phi(\widetilde{\mathbf{Z}}_{\Pi}) > \epsilon) + T\beta(a),$$

[7] The existence of a unique fixed point is guaranteed by Lemma 3.2 in (Bartlett et al, 2005).

where $\widetilde{\mathbf{Z}}_{\Pi}$ is an i.i.d. sample of size m from Π. By Theorem 3.3 of Bartlett et al (2005), 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 Bartlett et al (2005) can be extended to the setting of asymptotically stationary processes. Of course, these results also hold for stationary β-mixing processes since, as we pointed out in Section 6, these are just a special case of asymptotically stationary processes.

8 Unbounded Loss Functions

The learning guarantees that we have presented so far only hold for bounded loss functions. For a large variety of time series prediction problems, this assumption does not hold. We now demonstrate that the sub-sample selection technique of Proposition 2 enables us to extend the relative deviation bounds (Cortes et al, 2013; Vapnik, 1998) to the setting of asymptotically stationary processes, thereby providing guarantees for learning with unbounded losses in this scenario. In fact, since stationary mixing processes are asymptotically stationary, these results are the first generalization bounds for unbounded losses even in that simpler case.

The guarantees that we present are in terms of the expected number of dichotomies generated by a set $Q = \{(z,t) \mapsto \mathbf{1}_{\ell(h,z)\geq t} \colon h \in H, t \in \mathbb{R}\}$ over the sample $\mathbf{Z}_1^T$ that we denote by $\mathbb{S}_Q(\mathbf{Z}_1^T)$. We will also use the following notation for the αth moment of the loss function with respect to stationary distribution: $\mathcal{L}_{\Pi,\alpha}(h) = \mathbb{E}_{\Pi}[\ell(h,Z)^\alpha]$. Define

$$\Phi_{\tau,\alpha}(\mathbf{Z}_1^T) = \sup_h \left(\frac{\mathcal{L}_{\Pi}(h) - \frac{1}{T}\sum_{t=1}^T \ell(h,Z_t)}{(\mathcal{L}_{\Pi,\alpha} + \tau)^{1/\alpha}} \right).$$

Lemma 10 *Let $0 \leq \epsilon < 1$, $1 < \alpha \leq 2$, and $0 < \tau^{\frac{\alpha-1}{\alpha}} \leq \epsilon^{\frac{\alpha}{\alpha-1}}$. Let L be any (possibly unbounded) loss function and H any hypothesis set such that $\mathcal{L}_{\Pi,\alpha}(h) < \infty$ for all $h \in H$. Suppose that $T = ma$ for some $m, a > 0$. Then, for any $\epsilon > 0$, the following holds:*

$$\mathbb{P}(\Phi_{\tau,\alpha}(\mathbf{Z}_1^T) > \Gamma(\alpha,\epsilon)\epsilon) \leq a\mathbb{P}(\Phi_{\tau,\alpha}(\widetilde{\mathbf{Z}}_{\Pi}) > \Gamma(\alpha,\epsilon)\epsilon) + T\beta(a),$$

where $\widetilde{\mathbf{Z}}_{\Pi}$ is an i.i.d. sample of size m from Π and $\Gamma(\alpha,\epsilon) = \frac{\alpha-1}{\alpha}(1+\tau)^{\frac{1}{\alpha}} + \frac{1}{\alpha}(\frac{\alpha}{\alpha-1}^{\alpha-1})(1+(\frac{\alpha-1}{\alpha})^\alpha \tau^{\frac{1}{\alpha}})^{\frac{1}{\alpha}} \left[1 + (\frac{\alpha-1}{\alpha})^{\frac{\alpha-1}{\alpha}} \log(1/\epsilon)\right]^{\frac{\alpha-1}{\alpha}}$.

Proof We observe that the following holds:

$$\begin{aligned}
&\{\Phi_{\tau,\alpha}(\mathbf{Z}_1^T) > \Gamma(\alpha,\epsilon)\epsilon\} \\
&\quad= \left\{\exists h\colon \frac{1}{T}\sum_{t=1}^{T}(\mathcal{L}_\Pi(h) - \ell(h,Z_t)) > (\mathcal{L}_{\Pi,\alpha}+\tau)^{1/\alpha}\Gamma(\alpha,\epsilon)\epsilon\right\} \\
&\quad= \left\{\exists h\colon \frac{1}{am}\sum_{j=1}^{a}\sum_{t=0}^{m-1}(\mathcal{L}_\Pi(h) - \ell(h,Z_{ta+j})) > (\mathcal{L}_{\Pi,\alpha}+\tau)^{1/\alpha}\Gamma(\alpha,\epsilon)\epsilon\right\} \\
&\quad\subset \cup_{j=1}^{a}\left\{\exists h\colon \frac{1}{m}\sum_{t=0}^{m-1}(\mathcal{L}_\Pi(h) - \ell(h,Z_{ta+j})) > (\mathcal{L}_{\Pi,\alpha}+\tau)^{1/\alpha}\Gamma(\alpha,\epsilon)\epsilon\right\} \\
&\quad= \cup_{j=1}^{a}\{\Phi_{\tau,\alpha}(\mathbf{Z}^{(j)}) > \Gamma(\alpha,\epsilon)\epsilon\}.
\end{aligned}$$

Therefore, by Proposition 2 and the union bound the following result follows:

$$\begin{aligned}
\mathbb{P}(\Phi_{\tau,\alpha}(\mathbf{Z}_1^T) > \Gamma(\alpha,\epsilon)\epsilon) &\leq \sum_{j=1}^{a}\mathbb{P}(\Phi_{\tau,\alpha}(\mathbf{Z}^{(j)}) > \Gamma(\alpha,\epsilon)\epsilon) \\
&\leq a\mathbb{P}(\Phi_{\tau,\alpha}(\widetilde{\mathbf{Z}}_\Pi) > \Gamma(\alpha,\epsilon)\epsilon) + T\beta(a),
\end{aligned}$$

and this concludes the proof. □

Lemma 10, Corollary 13 and Corollary 14 in (Cortes et al, 2013) immediately yield the following learning guarantee for $\alpha = 2$.

Corollary 3 *With the assumptions of Lemma 10, for any $\delta > a(m-1)\beta(a)$, with probability $1-\delta$, the following holds for all hypothesis $h \in H$:*

$$\mathcal{L}_\Pi(h) \leq \sum_{t=1}^{T}\ell(h,Z_t) + 2\sqrt{\mathcal{L}_{\Pi,2}(h)}B_m\Gamma_0(2B_m)$$

where $\delta' = \delta - T\beta(a)$, $\Gamma_0(\epsilon) = \frac{1}{2} + \sqrt{1 + \frac{1}{2}\log\frac{1}{\epsilon}}$ and

$$B_m = \sqrt{\frac{2\log\mathbb{E}_\Pi[\mathbb{S}_Q(\mathbf{Z}_1^T)] + \log\frac{1}{\delta'}}{m}}.$$

This result generalizes i.i.d. learning guarantees with unbounded losses to the setting of non-i.i.d. data. Observe, that $\Gamma_0(2B_m)$ scales logarithmically with m and this bound admits $O(\log(m)/\sqrt{m})$ dependency. It is also possible to give learning guarantees in terms of higher order moments $\alpha > 2$.

Lemma 11 *Let $0 \leq \epsilon < 1$, $\alpha > 2$, and $0 < \tau \leq \epsilon^2$. Let L be any (possibly unbounded) loss function and H any hypothesis set such that $\mathcal{L}_{\Pi,\alpha}(h) < \infty$ for all $h \in H$. Suppose that $T = ma$ for some $m, a > 0$. Then, for any $\epsilon > 0$, the following holds:*

$$\mathbb{P}(\Phi_{\tau,\alpha}(\mathbf{Z}_1^T) > \Lambda(\alpha,\epsilon)\epsilon) \leq a\mathbb{P}(\Phi_{\tau,\alpha}(\widetilde{\mathbf{Z}}_\Pi) > \Lambda(\alpha,\epsilon)\epsilon) + T\beta(a),$$

where $\widetilde{\mathbf{Z}}_\Pi$ is an i.i.d. sample of size m from Π and $\Lambda(\alpha,\epsilon) = 2^{-2/\alpha}(\frac{\alpha}{\alpha-2})^{\frac{\alpha-1}{\alpha}} + \frac{\alpha}{\alpha-1}\tau^{\frac{\alpha-2}{2\alpha}}$.

Finally, it is also possible to extend the guarantees for the ERM algorithm with unbounded losses given for i.i.d. data in (Liang et al, 2015; Mendelson, 2014, 2015) to the setting of asymptotically stationary processes using our sub-sample selection technique.

9 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 by minimizing an upper bound on the discrepancy that can be computed from the data (Mansour et al, 2009; Kifer et al, 2004). 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 asymptotically stationary processes. In doing so, we have introduced a new sub-sample selection technique that can be of independent interest. We also 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 Bartlett et al (2005). Finally, we also provide the first learning guarantees for unbounded losses in the setting of non-i.i.d. data.

Acknowledgements We thank Marius Kloft and Andrés Muñoz Medina for discussions about topics related to this research. This work was partly funded by the NSF awards IIS-1117591 and CCF-1535987, a Google Research Award, and the National Science and Engineering Research Council of Canada PGS D3 award.

References

Agarwal A, Duchi J (2013) The generalization ability of online algorithms for dependent data. Information Theory, IEEE Transactions on 59(1):573–587

Alquier P, Wintenberger O (2010) Model selection for weakly dependent time series forecasting. Tech. Rep. 2010-39, Centre de Recherche en Economie et Statistique

Alquier P, Li X, Wintenberger O (2014) Prediction of time series by statistical learning: general losses and fast rates. Dependence Modelling 1:65–93

Bartlett PL, Bousquet O, Mendelson S (2005) Local rademacher complexities. Ann Statist 33(4):1497–1537

Barve RD, Long PM (1996) On the complexity of learning from drifting distributions. In: Proceedings of the Ninth Annual Conference on Computational Learning Theory, COLT '96

Bernstein S (1927) Sur l'extension du thorme limite du calcul des probabilits aux sommes de quantits dpendantes. Mathematische Annalen 97(1):1–59

Berti P, Rigo P (1997) A Glivenko-Cantelli theorem for exchangeable random variables. Statistics & Probability Letters 32(4):385 – 391

Cortes C, Greenberg S, Mohri M (2013) Relative deviation learning bounds and generalization with unbounded loss functions. CoRR abs/1310.5796

Doukhan P (1994) Mixing : properties and examples. Lecture notes in statistics, Springer-Verlag, New York

Dudley RM (2002) Real analysis and probability. Cambridge studies in advanced mathematics, Cambridge University Press, Cambridge

Eberlein E (1984) Weak convergence of partial sums of absolutely regular sequences. Statistics & Probability Letters 2(5):291 – 293

Hsu DJ, Kontorovich A, Szepesvari C (2015) Mixing time estimation in reversible markov chains from a single sample path. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in Neural Information Processing Systems 28, Curran Associates, Inc., pp 1459–1467

Kifer D, Ben-David S, Gehrke J (2004) Detecting change in data streams. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30, VLDB '04, pp 180–191

Koltchinskii V, Panchenko D (2000) Rademacher processes and bounding the risk of function learning. In: Gin E, Mason D, Wellner J (eds) High Dimensional Probability II, Progress in Probability, vol 47, Birkhuser Boston, pp 443–457

Kuznetsov V, Mohri M (2014) Generalization bounds for time series prediction with non-stationary processes. In: Auer P, Clark A, Zeugmann T, Zilles S (eds) Algorithmic Learning Theory, Lecture Notes in Computer Science, vol 8776, Springer International Publishing, pp 260–274

Kuznetsov V, Mohri M (2015) Learning theory and algorithms for forecasting non-stationary time series. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in Neural Information Processing Systems 28, Curran Associates, Inc., pp 541–549

Liang T, Rakhlin A, Sridharan K (2015) Learning with square loss: Localization through offset rademacher complexity. In: Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015, pp 1260–1285

Mansour Y, Mohri M, Rostamizadeh A (2009) Domain adaptation with multiple sources. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds) Advances in Neural Information Processing Systems 21, Curran Associates, Inc., pp 1041–1048

McDiarmid C (1989) On the method of bounded differences, Cambridge University Press, pp 148–188

Meir R (2000) Nonparametric time series prediction through adaptive model selection. Machine Learning pp 5–34

Mendelson S (2014) Learning without concentration. In: Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014, pp 25–39

Mendelson S (2015) Learning without concentration. J ACM 62(3):21

Mohri M, Muñoz AM (2012) New analysis and algorithm for learning with drifting distributions. In: Bshouty N, Stoltz G, Vayatis N, Zeugmann T (eds) Algorithmic Learning Theory, Lecture Notes in Computer Science, vol 7568, pp 124–138

Mohri M, Rostamizadeh A (2009) Rademacher complexity bounds for non-i.i.d. processes. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds) Advances in Neural Information Processing Systems 21, Curran Associates, Inc., pp 1097–1104

Mohri M, Rostamizadeh A (2010) Stability bounds for stationary φ-mixing and β-mixing processes. Journal of Machine Learning Research 11:789–814

Mohri M, Rostamizadeh A, Talwalkar A (2012) Foundations of Machine Learning. The MIT Press

De la Peña VH, Giné E (1999) Decoupling : from dependence to independence : randomly stopped processes, U-statistics and processes, martingales and beyond. Probability and its applications, Springer, New York

Pestov V (2010) Predictive PAC learnability: A paradigm for learning from exchangeable input data. In: Proceedings of the 2010 IEEE International Confer-

ence on Granular Computing, GRC '10, pp 387–391
Rakhlin A, Sridharan K, Tewari A (2010) Online learning: Random averages, combinatorial parameters, and learnability. In: Lafferty J, Williams C, Shawe-Taylor J, Zemel R, Culotta A (eds) Advances in Neural Information Processing Systems 23, Curran Associates, Inc., pp 1984–1992
Rakhlin A, Sridharan K, Tewari A (2011a) Online learning: Beyond regret. In: COLT 2011 - The 24th Annual Conference on Learning Theory, pp 559–594
Rakhlin A, Sridharan K, Tewari A (2011b) Online learning: Stochastic, constrained, and smoothed adversaries. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger K (eds) Advances in Neural Information Processing Systems 24, Curran Associates, Inc., pp 1764–1772
Rakhlin A, Sridharan K, Tewari A (2015) Sequential complexities and uniform martingale laws of large numbers. Probability Theory and Related Fields 161(1-2):111–153
Ralaivola L, Szafranski M, Stempfel G (2010) Chromatic pac-bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes. Journal of Machine Learning Research 11:1927–1956
Shalizi C, Kontorovich A (2013) Predictive PAC learning and process decompositions. In: Burges C, Bottou L, Welling M, Ghahramani Z, Weinberger K (eds) Advances in Neural Information Processing Systems 26, Curran Associates, Inc., pp 1619–1627
Steinwart I, Christmann A (2009) Fast learning from non-i.i.d. observations. In: Bengio Y, Schuurmans D, Lafferty J, Williams C, Culotta A (eds) Advances in Neural Information Processing Systems 22, Curran Associates, Inc., pp 1768–1776
Vapnik V (1998) Statistical learning theory. Wiley
Volkonskii V, Rozanov Y (1959) Some limit theorems for random functions. i. Theory of Probability & Its Applications 4(2):178–197
Yu B (1994) Rates of convergence for empirical processes of stationary mixing sequences. The Annals of Probability 22(1):94–116

A Proofs

Lemma 12 *Let $\mathbf{Y}$ be a weakly stationary processes, L be a squared loss function and $H = \{\mathbf{Y}^t_{t-q+1} \mapsto \mathbf{w} \cdot \mathbf{Y}^t_{t-q+1} \colon w \in \mathbb{R}^q\}$. Then $d(t_1, t_2) = 0$ for all t_1, t_2.*

Proof Observe that for any t_1 we can write

$$\mathbb{E}\left[\left(\mathbf{w} \cdot \mathbf{Y}^t_{t_1-q+1} - Y_{t_1+s}\right)^2\right] = \mathbb{E}\left[\left(\mathbf{w} \cdot \mathbf{Y}^t_{t_1-q+1}\right)^2\right] + \mathbb{E}[Y^2_{t_1+s}] - 2\,\mathbb{E}\left[\left(\mathbf{w} \cdot \mathbf{Y}^t_{t_1-q+1}\right) Y_{t_1+s}\right].$$

The first term on the right-hand side can be written as

$$\sum_{j,i=1}^{q} w_j w_i \,\mathbb{E}[Y_{t_1-i+1} Y_{t_1-j+1}] = \sum_{j,i=1}^{q} w_j w_i f(i-j)$$

for some function f, since $\mathbf{Y}$ is weakly stationary. Similarly we can write the last term as

$$\sum_{j}^{q} w_j f(s+j-1)$$

and the second term is $f(0)$. Therefore, we have that

$$\mathbb{E}\left[\left(\mathbf{w}\cdot\mathbf{Y}_{t_1-q+1}^{t}-Y_{t_1+s}\right)^2\right]=\sum_{j,i=1}^{q}w_jw_if(i-j)+f(0)-2\sum_{j}^{q}w_jf(s+j-1).$$

Observe that the right-hand side in the last equation is independent of t_1. This implies that $\mathcal{L}_{t_1}(h)=\mathcal{L}_{t_2}(h)$ for all t_1,t_2 and all $h\in H$, concluding the proof that $\bar{d}(t_1,t_2)=0$. □

B Review of Sequential Rademacher Complexity

One of the main ingredients for our generalization bounds in Section 5 is so called sequential Rademacher complexity originally introduced in (Rakhlin et al, 2010). Let $\mathcal{G}$ be a set of functions from $\mathcal{Z}$ to $\mathbb{R}$. Sequential Rademacher complexity of a function class $\mathcal{G}$ is defined to be

$$\mathfrak{R}_T^{seq}(\mathcal{G})=\frac{1}{T}\sup_{\mathbf{z}\in\mathcal{Z}}\mathbb{E}\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{T}\epsilon_t g(z_t(\epsilon))\right], \tag{15}$$

where supremum is taken over all complete binary trees of depth T with values in $\mathcal{Z}$ and ϵ is a sequence of Rademacher random variables. For our purposes we adopt the following definition of a complete binary tree. A $\mathcal{Z}$-valued complete binary tree $\mathbf{z}$ a sequence $(z_1,\ldots,z_T)$ where $z_t\colon\{\pm1\}^{t-1}\to\mathcal{Z}$. The reader should think of the root z_1 as some constant in $\mathcal{Z}$. The left child of the root is $z_2(-1)$ and the right child is $z_2(1)$. A path in the tree is $\epsilon=(\epsilon_1,\ldots,\epsilon_{T-1})$. To simplify the notation we will write $v_t(\epsilon)$ instead of $z_t(\epsilon_1,\ldots,\epsilon_{t-1})$. The following symmetrization result from (Rakhlin et al, 2015) is needed in the proof of Lemma 8.

Theorem 6 (Theorem 2 in (Rakhlin et al, 2015)) *The following bound holds*

$$\frac{1}{T}\,\mathbb{E}\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{T}\Big(\mathbb{E}\left[g(Z_{t+s})|\mathbf{Z}_1^t\right]-g(Z_{t+s})\Big)\right]\le 2\mathfrak{R}_T^{seq}(\mathcal{G}).$$

Proof The proof of this result is given in (Rakhlin et al, 2015) for the case $s=1$. We will now demonstrate that the same proof is valid for an arbitrary s. Let $\{Z'_t\}$ be a decoupled tangent sequence to $\{Z_t\}$. That is, Z'_{t+s} is drawn from $\mathbf{P}_{t+s}(\cdot|\mathbf{Z}_1^t)$ independently of $\mathbf{Z}_{t+1}^{\infty}$.[8] We will carry out the formal construction of this sequence at the end of this proof and in the meantime we assume that such a sequence always exists. Observe that definition implies that

$$\mathbb{E}[g(Z_{t+s})|\mathbf{Z}_1^t]=\mathbb{E}[g(Z'_{t+s})|\mathbf{Z}_1^t]=\mathbb{E}[g(Z'_{t+s})|\mathbf{Z}_1^{T+s}]$$

and also we have that $g(Z_{t+s})=\mathbb{E}[g(Z_{t+s})|\mathbf{Z}_1^{T+s}]$. Following the argument from (Rakhlin et al, 2015), we have that

$$\begin{aligned}\mathbb{E}\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{T}\Big(\mathbb{E}[g(Z_{t+s})|\mathbf{Z}_1^t]-g(Z_{t+s})\Big)\right]&=\mathbb{E}\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{T}\mathbb{E}\left[g(Z'_{t+s})-g(Z_{t+s})|\mathbf{Z}_1^{T+s}\right]\right]\\&\le\mathbb{E}\left[\sup_{g\in\mathcal{G}}\sum_{t=1}^{T}\big(g(Z'_{t+s})-g(Z_{t+s})\big)\right],\end{aligned}$$

where the inequality is a consequence of the linearity of expectation and Jensen's inequality. The next step in the proof of Rakhlin et al (2015) is to appeal to Lemma 18. Since Lemma 18 in (Rakhlin et al, 2015) is stated in terms of decoupled tangent sequences with $s=1$, we repeat the argument here for $s>1$.

[8] Note that the regular conditional law $\mathbf{P}_{t+s}(\cdot|\mathbf{Z}_1^t)$ exists provided $\mathcal{Z}$ is a Polish space (Dudley, 2002).

Observe that since Z_{t+s} and Z'_{t+s} are independent and identically distributed given $\mathbf{Z}_1^T$, if $\mathbb{E}_T$ denotes expectation with respect to Z'_{T+s}, Z_{T+s}, we must have that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{T}\Big[\sup_{g\in\mathcal{G}}\Big(\sum_{t=1}^{T-1}(g(Z'_{t+s})-g(Z_{t+s}))+(g(Z'_{T+s})-g(Z_{T+s}))\Big)|\mathbf{Z}_1^T,{\mathbf{Z}'}_1^T\Big]\\
&\quad=\mathop{\mathbb{E}}_{T}\Big[\sup_{g\in\mathcal{G}}\Big(\sum_{t=1}^{T-1}(g(Z'_{t+s})-g(Z_{t+s}))-(g(Z'_{T+s})-g(Z_{T+s}))\Big)|\mathbf{Z}_1^T,{\mathbf{Z}'}_1^T\Big]\\
&\quad=\mathop{\mathbb{E}}_{\epsilon_T}\mathop{\mathbb{E}}_{T}\Big[\sup_{g\in\mathcal{G}}\Big(\sum_{t=1}^{T-1}(g(Z'_{t+s})-g(Z_{t+s}))+\epsilon_T(g(Z'_{T+s})-g(Z_{T+s}))\Big)|\mathbf{Z}_1^T,{\mathbf{Z}'}_1^T\Big]\\
&\quad=\mathop{\mathbb{E}}_{T}\mathop{\mathbb{E}}_{\epsilon_T}\Big[\sup_{g\in\mathcal{G}}\Big(\sum_{t=1}^{T-1}(g(Z'_{t+s})-g(Z_{t+s}))+\epsilon_T(g(Z'_{T+s})-g(Z_{T+s}))\Big)|\mathbf{Z}_1^T,{\mathbf{Z}'}_1^T\Big]\\
&\quad\leq\sup_{z_{T+s},z'_{T+s}\in\mathcal{Z}^2}\mathop{\mathbb{E}}_{\epsilon_T}\Big[\sup_{g\in\mathcal{G}}\Big(\sum_{t=1}^{T-1}(g(Z'_{t+s})-g(Z_{t+s}))+\epsilon_T(g(z'_{T+s})-g(z_{T+s}))\Big)\Big].
\end{aligned}$$

Iterating the above inequality and using the tower property of the conditional expectation as in (Rakhlin et al, 2015), we obtain

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{g\in\mathcal{G}}\sum_{t=1}^{T}\Big(\mathbb{E}[g(Z_{t+s})|\mathbf{Z}_1^t]-g(Z_{t+s})\Big)\Big]\\
&\quad\leq\sup_{z_{1+s},z'_{1+s}}\mathop{\mathbb{E}}_{\epsilon_1}\ldots\sup_{z_{T+s},z'_{T+s}}\mathop{\mathbb{E}}_{\epsilon_T}\Big[\sup_{g\in\mathcal{G}}\Big(\sum_{t=1}^{T}\epsilon_t(g(z'_{t+s})-g(z_{t+s}))\Big)\Big]\\
&\quad\leq 2\sup_{z_{1+s}}\mathop{\mathbb{E}}_{\epsilon_1}\ldots\sup_{z_{T+s}}\mathop{\mathbb{E}}_{\epsilon_T}\Big[\sup_{g\in\mathcal{G}}\Big(\sum_{t=1}^{T}\epsilon_t g(z_{t+s})\Big)\Big].
\end{aligned}$$

The last upper bound precisely matches Equation (14) from (Rakhlin et al, 2015) (up to re-parametrization) and the rest of the argument is the same.

To complete the proof we show that the decoupled tangent sequence always exist. Existence of such a sequence in case $s=1$ is well-known (see for example (De la Peña and Giné, 1999)). We show that the standard construction also works for an arbitrary s. If $\mathbf{Z}$ is a sequence of random variables defined on the probability triple $(\Omega,\Sigma,\mathbf{P})$ and taking values in $(\mathcal{Z},\mathcal{B})$, where $\mathcal{Z}$ is a Polish space and $\mathcal{B}$ is its Borel σ-algebra. Then consider a measure space a measure space $(\Omega\times\mathcal{Z}^{\mathbb{N}},\Sigma\times\mathcal{B}^{\mathbb{N}})$. Define a probability measure $\widehat{\mathbf{P}}$ on this extended measure space by

$$\widehat{\mathbf{P}}(A\times B)=\mathop{\mathbb{E}}_{\mathbf{P}}[\otimes_{t=1}^{\infty}\mathbf{P}_{t+s}(B|\mathbf{Z}_1^t)\mathbf{1}_A]=\int_A\otimes_{t=1}^{\infty}\mathbf{P}_{t+s}(B|\mathbf{Z}_1^t)(w)d\mathbf{P}(w)$$

With out loss of generality, we may assume that Z_t is defined on the extended measure space by $Z_t(w,\mathbf{z})=Z_t(w)$ since $Z_t(w,\mathbf{z})$ and Z_t have the same finite dimensional distributions. We define $Z'_t(w,\mathbf{z})=z_t$. From this construction, it follows that Z_t and Z'_T are decoupled tangent sequences and the proof is complete. □

C McDiarmid's inequality for dependent random variables

One of the main ingredients of our bounds in Section 5 is a version of McDiarmid's inequality for dependent random variables from (McDiarmid, 1989). For convenience of our reader, we state this result in the next theorem.

Theorem 7 (Corollary 6.10 in (McDiarmid, 1989)) *Let $Z_1,\ldots,Z_T$ be $\mathcal{Z}$-valued random variables and $\Phi\colon\mathcal{Z}^T\to\mathbb{R}$ be a Borel-measurable function such that there exist non-negative constants $c_1,\ldots,c_T$ satisfying*

$$|\,\mathbb{E}[\Phi(\mathcal{Z}_1^T)|z_1,\ldots,z_t]-\mathbb{E}[\Phi(\mathcal{Z}_1^T)|z_1,\ldots,z'_t]|\leq c_t$$

for all $z_1, \ldots, z_t, z_t' \in \mathcal{Z}$*. Then for any* $\epsilon > 0$ *the following inequality holds*

$$\mathbb{P}(\Phi(\mathcal{Z}_1^T) - \mathbb{E}\,\Phi(\mathcal{Z}_1^T) \geq \epsilon) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{t=1}^T c_t^2}\right).$$