Generalization for slowly mixing processes

Andreas Maurer
Istituto Italiano di Tecnologia, CSML, 16163 Genoa, Italy
am@andreas-maurer.eu

Abstract

A bound uniform over various loss-classes is given for data generated by stationary and φ-mixing processes, where the mixing time (the time needed to obtain approximate independence) enters the sample complexity only in an additive way. For slowly mixing processes this can be a considerable advantage over results with multiplicative dependence on the mixing time. The admissible loss-classes include functions with prescribed Lipschitz norms or smoothness parameters. The bound can also be applied to be uniform over unconstrained loss-classes, where it depends on local Lipschitz properties of the function in question.

1 Introduction

A key problem in learning theory is to give performance guarantees on new, yet unseen data for hypotheses which are selected on the basis of their performance on a finite number of observations. To bound the expected loss in terms of the observed average loss with high probability uniformly over various classes of constrained hypotheses many techniques have been developed. These methods are quite effective, when the observations are independent, but there are significant difficulties when they become dependent, as happens for many stochastic processes, for example in the study of dynamical systems. A popular approach assumes the process to be stationary and mixing, so that future observations are nearly independent of a sufficiently distant past.

If the X_i are observations becoming approximately independent after τ increments of time, then $(X_1, X_{1+\tau}, X_{1+2\tau}, ..., X_{1+n\tau})$ can be treated as a vector of n independent observations, to which the law of large numbers can be applied, modulo a correction term dependent on τ. An approach based on this idea using nearly independent data blocks has been introduced by [22] and since been used in various forms by many authors ([15], [16], [20], [17], [1], and others) to port established techniques and results from the independent to the dependent setting.

There are two problems with this approach. The more philosophical one is, that, while mixing appears to be a sensible assumption in many situations, it is difficult to estimate its quantitative properties from a single sample path, although some significant progress has been made recently for Markov chains ([11], [21]).

The other more practical limitation is that to obtain the same bound on the estimation error as for independent data, the number of necessary observations is multiplied with the mixing time τ. This is a major problem when the process mixes very slowly. On the other hand it seems difficult to obtain general results, when the mixing assumption is abandoned altogether, although there are results without mixing in the more specialized settings of linear and generalized linear dynamical systems ([19], [10]).

This work presents a data-dependent generalization bound for φ-mixing processes with stationary distribution π.The π-expected risk of hypotheses and the probability of excessive losses are bounded by an empirical functional with high probability uniform over the class of hypotheses.

One advantage of the approach is that the mixing time enters only additively in the required number of observations. For slowly mixing processes this can reduce the sample complexity by an order of magnitude.

Another advantage is that it allows for rather large function classes, such as the classes of all functions with prescribed Lipschitz norm or γ-smoothness. Constraints on the Lipschitz norm have been used in generalization bounds for deep neural networks ([3]). Because of its small constants and strong data-dependence our bound may be competitive even in the iid setting, at least for favourable data distributions. It will also be shown below, that our bound can be uniform over completely unconstrained loss classes, where it depends only on local properties of the functions at the sample path.

The price paid is that we abandon the average loss on the observations and replace it by a maximum. This obvious disadvantage is somewhat alleviated by the possibility to allow a small number of outliers, so the strict maximum can be replaced by the maximal loss on most of the observations. Another disadvantage is, that we require the stronger φ-mixing assumption, while for most previous results the weaker β- or α-mixing assumptions suffice.

2 Notation and preliminaries

We use capital letters for random variables, bold letters for vectors, and the set $\{1, ..., m\}$ is denoted $[m]$. The cardinality, complement and indicator function of a set A are denoted $|A|$, A^c and $\mathbf{1}A$ respectively. For a real-valued function f on a set A the supremum of its values is denoted $\|f\|_\infty$.

Throughout the following $(\mathcal{X}, \sigma)$ is a measurable space and $\mathbf{X} = (X_i)_{i\in\mathbb{N}}$ a stochastic process with values in $\mathcal{X}$. For $I \subseteq \mathbb{N}$, $\sigma(I)$ denotes the sigma-field generated by $(X_i)_{i\in I}$ and μ_I the corresponding joint marginal. The process is assumed to be stationary, so that $\forall I \subseteq \mathbb{Z}$, $i \in \mathbb{N}$, $\mu_I = \mu_{I+i}$, the stationary distribution being denoted $\pi = \mu_{\{0\}} = \mu_{\{k\}}$. It is called ergodic if for every $A \in \sigma$ with $\pi(A) > 0$ we have $\Pr\{\forall k \leq n, X_k \notin A\} \to 0$ as $n \to \infty$. The φ-mixing and α-mixing coefficients ([22], [6]) are defined for any $\tau \in \mathbb{N}$ as

$$\begin{aligned}\varphi_\tau &= \sup\{|\Pr(A|B) - \Pr A| : k \in \mathbb{Z}, A \in \sigma(\{k\}), B \in \sigma(\{i : i < k - \tau\})\},\\ \alpha_\tau &= \sup\{|\Pr(A \cap B) - \Pr A \Pr B| : k \in \mathbb{Z}, A \in \sigma(\{k\}), B \in \sigma(\{i : i < k - \tau\})\}.\end{aligned}$$

The process is called φ-mixing (α-mixing) if $\phi_\tau \to 0$ ($\alpha_\tau \to 0$) as $\tau \to \infty$.

A *loss class* $\mathcal{F}$ is a set of measurable functions $f : \mathcal{X} \to [0, \infty)$, where f is to be thought of as a hypothesis composed with a fixed loss function. Very often $\mathcal{X} = \mathcal{Z} \times \mathcal{Z}'$, where $\mathcal{Z}$ is a measurable space of "inputs", $\mathcal{Z}'$ is a space of "outputs", "covariates" or "labels", $\mathcal{H}$ is a set of functions $h : \mathcal{Z} \to \mathbb{R}$ and ℓ is a fixed loss function $\ell : \mathbb{R} \times \mathcal{Z}' \to [0, \infty)$. The loss class in question would then be the class of functions

$$\mathcal{F} = \{(z, z') \mapsto \ell(h(z), z') : h \in \mathcal{H}\}.$$

3 Gauge pairs and generalization

Definition 3.1. *Let $\mathcal{F}$ be a class of nonnegative loss functions on a space $\mathcal{X}$. A gauge pair for $\mathcal{F}$ is a pair (g, Φ) where g is a measurable function $g : \mathcal{X} \times \mathcal{X} \to [0, \infty]$ such that $g(x, y) = 0$ iff $x = y$, and Φ is a function $\Phi : \mathcal{F} \times \mathcal{X} \to [0, \infty]$ such that for all $x, y \in \mathcal{X}$ and $f \in \mathcal{F}$, $\Phi(f, \cdot)$ is measurable and*

$$f(y) \leq g(y, x) + \Phi(f, x).$$

Intuitively $g(y, x)$ should measure the extent to which members of $\mathcal{F}$ can generalize from an observed datum x to a yet unobserved datum y. It helps to think of g as a nondecreasing function of a metric, but g need not be symmetric.

An example. The simplest example is furnished by the class of L-Lipschitz functions on a metric space $(\mathcal{X}, d)$. This means that $f(y) - f(x) \leq Ld(y, x)$ for all $x, y \in \mathcal{X}$ and $f \in \mathcal{F}$. Adding $f(x)$ to the inequality shows that $(g : (y, x) \mapsto Ld(y, x), \Phi : (f, x) \mapsto f(x))$ is a gauge pair. In the most typical case, when $\mathcal{X} = \mathcal{Z} \times \mathcal{Z}'$, a product metric is used. In classification, when $\mathcal{Z}'$ is a discrete set of labels, a discrete metric is used on$\mathcal{Z}'$.

More concretely $\mathcal{Z}$ could be a subset of $\mathbb{R}^D$, $\mathcal{Z}' = \{-1, 1\}$, $\mathcal{H} : \mathcal{Z} \to \mathbb{R}$ a set of neural networks of (euclidean) Lipschitz norm at most L, and $\ell : (t, y) \in \mathbb{R} \times \mathcal{Z}' \mapsto \eta(yt) \in [0, \infty)$, where η is a hinge or logistic loss. This is a standard situation in binary classification. The loss class is $\mathcal{F} = \{(z, z') \mapsto \ell(\langle h, z\rangle, z') : h \in \mathcal{H}\}$. Define $d((y, y'), (x, x')) = \|y - x\| + B|y' - x'|$ and verify that $\ell(h(y), y') - \ell(h(x), x') \leq Ld((y, y'), (x, x'))$.

Before giving more examples of gauge pairs we state our main result. One part bounds the probability of excessive losses, the other part bounds the risk properly.

Theorem 3.2. *Let $\mathbf{X} = (X_i)_{i\in\mathbb{N}}$ be a stationary process with values in $\mathcal{X}$ and invariant distribution π, $\mathcal{F}$ a class of measurable functions $f : \mathcal{X} \to [0, \infty)$ and (g, Φ) a gauge pair for $\mathcal{F}$. Let $X \sim \pi$ be independent of $\mathbf{X}$, $\tau \in \mathbb{N}$, $n > \tau$, $S \subseteq [n - \tau]$ and $\delta > 0$. Let $a \in \{1, 2\}$.*

(i) If $a = 1$, for any $t > 0$, with probability at least $1 - \delta$ in the sample path $\mathbf{X}_1^n = (X_1, ..., X_n)$

$$\sup_{f\in\mathcal{F}} \Pr\left\{ f(X) > \max_{j\in S\cap[n-\tau]} \Phi(f, X_j) + t \mid \mathbf{X}_1^n \right\}$$
$$\leq \frac{a}{n-\tau} \sum_{k=\tau+1}^{n} \mathbf{1}\left\{ \min_{i\in S\cap[k-\tau]} g(X_k, X_i) > t \right\} + \varphi(\tau) + \sqrt{\frac{2\ln(1/\delta)}{n-\tau}}.$$

(ii) If $a = 1$, with $\|\mathcal{F}\|_\infty = \sup_{f\in\mathcal{F}, x\in\mathcal{X}} f(x)$ and $\|g\|_\infty = \sup_{x,y\in\mathcal{X}} g(y, x)$, then with probability at least $1 - \delta$ in the sample path

$$\sup_{f\in\mathcal{F}} \mathbb{E}[f(X)] - \max_{i\in S\cap[n-\tau]} \Phi(f, X_i)$$
$$\leq \frac{a}{n-\tau} \sum_{k=\tau+1}^{n} \min_{i\in S\cap[n-\tau]} g(X_k, X_i) + \|\mathcal{F}\|_\infty \varphi_\tau + \|g\|_\infty \sqrt{\frac{2\ln(1/\delta)}{n-\tau}}.$$

(iii) For $a = 2$ the term $\sqrt{2\ln(1/\delta)/(n-\tau)}$ in both bounds can be replaced by $e\ln(1/\delta)/(n-\tau)$.

Remark 1. Both (i) and (ii) have two data-dependent terms, the first of which depends on the choice of the function f. It is the term, which a learning algorithm should try to minimize. In the example above it depends only on the evaluation of f at the sample points, but in general it may depend on any local properties of f. In part (i), which limits the probability of excessive losses, the appearance of the maximum is perhaps more natural. In both (i) and (ii) the first term on the R.H.S. is expected to be the dominant term. The preferred choice is $a = 1$, and part (iii) may have less practical relevance.

Remark 2. The presence of the maximum on the L.H.S. is clearly an unpleasant feature. A union bound over the "good" set S can allow for a small fraction of errors. This problem will be addressed in the next section, where it is shown that a union bound over the "good" set S (introduced exclusively for this purpose) can allow for a small fraction of errors. The bounds are intended to target the *realizable* or *nearly realizable* case, when the underlying function is in $\mathcal{F}$, with small noise and very few outliers, which might be hidden in the complement of the set S. Note that many modern learning machines which implement Lipschitz functions achieve near zero training error.

Remark 3. The terms

$$G_t = \frac{2}{n-\tau} \sum_{k=\tau+1}^{n} \mathbf{1}\left\{ \min_{i\in S\cap[k-\tau]} g(X_k, X_i) > t \right\} \text{ or } G = \frac{2}{n-\tau} \sum_{k=\tau+1}^{n} \min_{i\in S\cap[k-\tau]} g(X_k, X_i)$$

can be computed from the data and interpreted as a complexity of the sample path relative to g. In Section 5 it will be proven, that, under an assumption of total boundedness of the support of π (defined relative to g) and ergodicity of the process, the terms converge to zero in probability. For sufficiently fast α-mixing they can be shown to converge to zero almost surely. This condition is sufficient, but not necessary. The key property is recurrence, which may be independent of mixing and does not require an approach to the stationary distribution. Section 5 concludes with an example of recurrence and arbitrarily slow mixing.

Remark 4. Setting $\tau = 1$ and $\varphi_\tau = 0$ gives a version for iid variables. This may be interesting in its own right in comparison with other bounds for function classes with bounded Lipschitz norm, for

which we take [3] as a prototypical example. We do not expect Theorem 3.2 to be competitive in the general iid setting, but the applicability to unconstrained function classes, as shown in Section 6, might be an advantage.

The proof of Theorem 3.2 (iii) requires the following tail bound for martingale difference sequences, which is established in Section A.1.

Lemma 3.3. *Let $R_1, ..., R_n$ be real random variables $0 \leq R_j \leq 1$ and let $\sigma_1 \subseteq \sigma_2 \subseteq ...\sigma_n$ be a filtration such that R_j is σ_j-measurable. Let $\hat{V} = \frac{1}{n}\sum_j R_j$, $V = \frac{1}{n}\sum_j \mathbb{E}[R_j|\sigma_{j-1}]$. Then*

$$\Pr\left\{V > 2\hat{V} + t\right\} \leq e^{-nt/e}$$

and equivalently, for $\delta > 0$,

$$\Pr\left\{V > 2\hat{V} + \frac{e\ln(1/\delta)}{n}\right\} \leq \delta.$$

Proof of Theorem 3.2. Let $f \in \mathcal{F}$. From stationarity we obtain for every k, $\tau < k \leq n$ and $u > 0$

$$\begin{aligned}\Pr\{f(X) > u\} &= \Pr_{X\sim\mu_k}\{f(X) > u\} \\ &\leq \Pr\left\{f(X_k) > u \mid (X_i)_{i\in[k-\tau]}\right\} + \varphi(\tau),\end{aligned}$$

where the inequality follows from the definition of the φ-mixing coefficients, which allows us to replace the independent variable X by the sample path observation X_k, conditioned on observations more than τ time increments in the past. But if $f(X_k) > u$, then by the definition of gauge pairs we must have $g(X_k, X_i) + \Phi(f, X_i) > u$, for every $i \in [k-\tau]$, or, equivalently, $\min_{i\in[k-\tau]} g(X_k, X_i) + \Phi(f, X_i) > u$, which certainly implies $\min_{i\in S\cap[k-\tau]} g(X_k, X_i) + \max_{j\in S\cap[n-\tau]} \Phi(f, X_j) > u$. Thus

$$\Pr\{f(X) > u\} \leq \Pr\left\{\min_{i\in S\cap[k-\tau]} g(X_k, X_i) + \max_{j\in S\cap[n-\tau]} \Phi(f, X_j) > u \mid (X_i)_{i\in[k-\tau]}\right\} + \varphi(\tau).$$

Averaging this inequality over all values of k, $\tau < k \leq n$, and a change of variables $t = u - \max_{j\in S\cap[n-\tau]} \Phi(f, X_j)$ gives

$$\begin{aligned}&\Pr\left\{f(X) > \max_{j\in S\cap[n-\tau]} \Phi(f, X_j) + t \mid \mathbf{X}_1^n\right\} \\ &\leq \frac{1}{n-\tau}\sum_{k=\tau+1}^{n} \Pr\left\{\min_{i\in S\cap[k-\tau]} g(X_k, X_i) > t \mid (X_i)_{i\in[k-\tau]}\right\} + \varphi(\tau).\end{aligned} \tag{1}$$

We first prove (i). Let σ_k be the σ-algebra generated by $(X_i)_{i\in[k]}$ and $R_k = \mathbf{1}\left\{\min_{i\in S\cap[k-\tau]} g(X_k, X_i) > t\right\}$, so that R_k is σ_k-measurable. Then $\mathbb{E}[R_k|\sigma_{k-1}] = \Pr\left\{\min_{i\in S\cap[k-\tau]} g(X_k, X_i) > t \mid (X_i)_{i\in[k-\tau]}\right\}$ and thus $\mathbb{E}[R_k|\sigma_{k-1}] - R_k$ is a Martingale difference sequence with values in $[-1, 1]$ and by the Hoeffding-Azuma inequality ([14])

$$\begin{aligned}&\frac{1}{n-\tau}\sum_{k=\tau+1}^{n} \Pr\left\{\min_{i\in S\cap[k-\tau]} g(X_k, X_i) > t \mid (X_i)_{i\in[k-\tau]}\right\} \\ &\leq \frac{1}{n-\tau}\sum_{k=\tau+1}^{n} \mathbf{1}\left\{\min_{i\in S\cap[k-\tau]} g(X_k, X_i) > t\right\} + \sqrt{\frac{2\ln(1/\delta)}{n-\tau}}.\end{aligned}$$

Substitute in the right hand side of (1). As the the right hand side is independent of f, we can take the supremum over f on the left hand side to complete the proof of (i).

(ii) We use integration by parts and integrate the left hand side of (1) from zero to $\|\mathcal{F}\|_\infty$. This gives

$$\begin{aligned}
&\mathbb{E}\left[f\left(X\right)\right] - \max_{j\in S\cap[n-\tau]} \Phi\left(f, X_j\right) \\
&\leq \frac{1}{n-\tau}\sum_{k=\tau+1}^{n}\int_0^{\|\mathcal{F}\|_\infty} \mathbb{E}\left[\mathbf{1}\left\{\min_{i\in S\cap[k-\tau]} g\left(X_k, X_i\right) > t\right\} \mid \left(X_i\right)_{i\in[k-\tau]}\right] dt + \|\mathcal{F}\|_\infty \varphi\left(\tau\right) \\
&= \frac{1}{n-\tau}\sum_{k=\tau+1}^{n} \mathbb{E}\left[\min_{i\in S\cap[k-\tau]} g\left(X_k, X_i\right) \mid \left(X_i\right)_{i\in[k-\tau]}\right] + \|\mathcal{F}\|_\infty \varphi\left(\tau\right).
\end{aligned}$$

Now we use the Hoeffding-Azuma inequality just as in part (i). Again we can take the supremum on the left hand side. This completes the proof of (ii).

To prove (iii) use Lemma 3.3 instead of the Hoeffding-Azuma inequality ([14]). □

4 Relaxation of the maximal loss

To allow a fixed fraction of excess errors we use a union bound over all possible sets S of fixed cardinality as in Theorem 3.2, where the complement $[n-\tau]\setminus S = S^c$ is the set of "bad data points" or outliers. Now let $\alpha = (n-\tau-|S|)/(n-\tau)$ be the allowed fraction of excess errors. The union bound replaces the term $\ln(1/\delta)$ by $\ln\left(\binom{n-\tau}{m}/\delta\right)$. From Stirling's approximation we can obtain for $N\in\mathbb{N}$ the bound

$$\ln\binom{N}{\alpha N} \leq NH\left(\alpha\right) + \text{Rest}\left(N,\alpha\right),$$

where $H\left(\alpha\right) = \alpha\ln\frac{1}{\alpha} + (1-\alpha)\ln\frac{1}{1-\alpha}$ is the entropy of an α-Bernoulli variable and

$$\text{Rest}\left(N,\alpha\right) = -\ln\left(\sqrt{2\pi\alpha\left(1-\alpha\right)N}\right) + \frac{1}{12N} \leq \begin{cases} 0 & \text{if } 2\pi N \geq \frac{1}{\alpha(1-\alpha)} \\ \frac{\ln(\pi N/2)}{2} & \text{otherwise} \end{cases}.$$

This leads to the following version of Theorem 3.2 (i), where for simplicity we give only the second conclusion.

Corollary 4.1. *Under the condition of Theorem 3.2 let $\alpha\in[0,1]$ be such that $\alpha(n-\tau)\in\mathbb{N}$. Then with probability at least $1-\delta$ in the sample path we have for every $S\subseteq[n-\tau]$ of cardinality $\alpha(n-\tau)$ and every $f\in\mathcal{F}$ that*

$$\begin{aligned}
&\mathbb{E}_\pi\left[f\left(X\right)\right] - \max_{i\in[n-\tau]\cap S}\Phi\left(f, X_i\right) \\
&\leq \frac{2}{n-\tau}\sum_{k=\tau+1}^{n}\min_{i\in[k-\tau]\cap S} g\left(X_k, X_i\right) + \|\mathcal{F}\|_\infty\varphi_\tau + e\|g\|_\infty\left(H\left(\alpha\right) + \frac{\textit{Rest}\left(n-\tau,\alpha\right) + \ln\left(1/\delta\right)}{n-\tau}\right).
\end{aligned}$$

Remark. Ignoring the Rest-term, which is at most of order $\ln(n)/n$, there is an additional penalty $e\|g\|_\infty H\left(\alpha\right)$ depending on the error fraction α. This can be interpreted as an additional empirical error term. Since $H\left(\alpha\right)\to 0$ as $\alpha\to 0$, the bound tolerates a small number of excess errors. On the other hand $H\left(\alpha\right)/\alpha\to\infty$ logarithmically as $\alpha\to 0$, so the penalty is certainly exaggerated relative to a conventional empirical error term, which would simply be α.

5 The complexity of the sample path

The dominant term on the right hand side of the bounds of Theorem 3.2 is likely to be the complexity term

$$G\left(\mathbf{X}, n, \tau, g\right) = \frac{1}{n-\tau}\sum_{k=\tau+1}^{n}\min_{i\in[k-\tau]} g\left(X_k, X_i\right),$$

where for simplicity we omit the set S and concentrate on the term as in appears in part (ii) of the theorem. It is a merit of Theorem 3.2 that we can observe this quantity directly and thus take

advantage of favourable situations. To get some idea of what characterizes these favourable situations it is nevertheless interesting to study the behavior of G in general, although our quantitative worst-case guarantees are disappointing.

Some standard concepts of the theory of metric spaces, such as diameters and covering numbers, extend to the case, when the metric is replaced by the function g.

Definition 5.1. *The g-diameter $diam_g(F)$ is $\sup_{x,y\in F} g(y,x)$. For $A \subseteq \mathcal{X}$ and $\epsilon > 0$ we write*

$$N(A,g,\epsilon) = \min\left\{ N : \exists C_1,...,C_N,\ diam_g(C_i) \leq \epsilon \text{ and } A \subseteq \bigcup_{j\in[N]} C_j \right\}.$$

$A \subseteq \mathcal{X}$ is g-totally bounded if $N(A,g,\epsilon) < \infty$ for every $\epsilon > 0$.

This definitions are intuitive if one thinks of g as an increasing function of the metric. It must still be kept in mind that g may not be symmetric nor obey the triangle inequality. The proof of the following theorem is given in Section A.3.

Theorem 5.2. *Let $\mathbf{X} = (X_i)_{i\in\mathbb{N}}$ be a stationary process and assume the support of π to be g-totally bounded.*

(i) If $\mathbf{X}$ is ergodic, then for any $\tau \in \mathbb{N}$, we have $G(\mathbf{X},n,\tau,g) \to 0$ in probability as $n \to \infty$

(ii) If there exists $q > 1$ and $A > 0$ such that the α-mixing coefficients satisfy $\alpha_\tau \leq A\tau^{-q}$ for all $\tau \in \mathbb{N}$, then for $p \in (2/(1+q),1)$ we have $G(\mathbf{X},n,\lceil n^p\rceil,g) \to 0$ almost surely.

(iii) In general we have

$$\Pr\{G(\mathbf{X},n,\tau,g) > t\} \leq \frac{N(supp(\pi),g,t/2)}{e(\lfloor nt/(2\tau)\rfloor - 1)} + \left\lceil \frac{nt}{2\tau} \right\rceil \alpha_\tau.$$

While part (i) is encouraging, the worst case bound (iii) is very weak for several reasons.

1. It scales with the covering number of the support of π. This behavior persists in the iid case, when $\tau = 0$ and $\alpha_\tau = 0$. Let us accept this scaling at face value. We can expect the bound of Theorem 3.2 to be strong only if the support of the invariant distribution is a small and essentially low dimensional object, even though it may be embedded in a complicated way in some high or even infinite dimensional ambient space.

2. It depends strongly on the mixing properties of the process. In particular it scales with τ/n, so as to exhibit a multiplicative scaling of the sample complexity with the mixing time, and to refute one of the principal claims made about this paper.

On the one hand these weaknesses highlight the benefit of observing G directly. More importantly (iii) remains a worst case bound, and the multiplicative scaling with τ is not generic. The decay of G depends on recurrence rather than mixing, and mixing is just the only way to give general quantitative bounds on recurrence. M. Kac [13] has already shown that for an ergodic process the expected recurrence time of a set A is $1/\pi(A)$, but this appears to be the only moment of the recurrence time, which can be controlled without mixing. For α and φ-mixing processes Chazottes [8] gives rapidly decreasing tails of the recurrence time, but these would lead to similar bounds as in the proof of (ii) above.

While mixing implies ergodicity and recurrence, the converse does not hold. A simple example is the deterministic unit rotation on the N-cycle, the transition matrix being

$$P(i,j) = \delta_{i,(j+1) \mod N}.$$

This Markov chain is ergodic and not mixing, but the recurrence time is N and obviously $G(\mathbf{X},n,N,g) = 0$ for all $n > N$ with g being the discrete metric $g(y,x) := 1$ iff $y \neq x$ and $g(x,x) := 0$. Of course Theorem 3.2 does not apply, but if we add some randomness, say

$$P(i,j) = (1-p)\,\delta_{i,(j+1) \mod N} + \frac{p}{N}, \text{ for } p > 0$$

the process becomes exponentially mixing, with spectral gap p, relaxation time $\tau_{rel} = p^{-1}$ and mixing time $\tau \geq p^{-1} \ln(1/\epsilon)$ to have distance ϵ from stationarity. This gives an estimate of the mixing coefficients $\alpha_\tau \geq \exp(-\tau p)$. The probability that the process visits all states in N steps is $(1-p)^N$, and the probability that it hasn't visited all states in τ steps is bounded by $\left(1-(1-p)^N\right)^{\tau/N}$. As $p \to 0$ the mixing time diverges, but the recurrence times converges to N. There are 2^N discrete Lipschitz functions, so the dominant term in the classical bounds (assuming the realizable case) would scale as τN while G scales as N.

Clearly a similar phenomenon is expected with irrational rotations on the circle and more general for any quasiperiodic motion, which means that the infinite trajectories are dense on the support of the invariant measure. Arnold and Avez [2] have shown that classical dynamical systems may be periodic, quasiperiodic or chaotic, and that the last two cases are generic. Adding a perturbation can make a quasiperiodic system mixing, but the mixing can be made arbitrarily slow.

6 Gauge pairs and unconstrained function classes

In Section 3 there was the more concrete example of Lipschitz classes, where the function g was essentially the metric. In this section we briefly discuss smooth classes and then show, that generalization bounds are possible even for completely unconstrained function classes.

Smooth functions. A real-valued, differentiable function on an open subset $\mathcal{O}$ of a Hilbert space is called γ-smooth if $\gamma > 0$ and for all $x, y \in \mathcal{X}$

$$\|f'(y) - f'(x)\| \leq \gamma \|y - x\|.$$

γ-smoothness plays a role in non-convex optimization because of the descent-lemma, giving a justification to the method of gradient descent [7]. Nonnegative γ-smooth functions satisfy special inequalities as in the following lemma, with proof Section A.2.

Lemma 6.1. *Let f be a nonnegative, differentiable, γ-smooth function on a convex open subset $\mathcal{O}$ of a Hilbert space and $\lambda > 0$. Then for any $x, y \in \mathcal{O}$*

$$f(y) \leq \left(1 + \lambda^{-1}\right) f(x) + (1+\lambda) \frac{\gamma}{2} \|y - x\|^2.$$

If $f(x) = 0$ then $f(y) \leq \frac{\gamma}{2} \|y - x\|^2$.

So $g : (y, x) \mapsto (1+\lambda) \frac{\gamma}{2} \|y - x\|^2$ and $(f, x) \mapsto \left(1 + \lambda^{-1}\right) f(x)$ make a gauge pair for the class of γ-smooth functions on $\mathcal{O}$, to which Theorem 3.2 could be applied.

Intuitively the passage from the metric $g \approx \|y - x\|$ to its square $g \approx \|y - x\|^2$ has positive consequences for the data-complexity term G. This is clear in terms of covering numbers, as they appear in Theorem 5.2 in the Section 5. If $N(\operatorname{supp}(\pi), g, t)$ is the number of sets C with $\sup_{x,y \in C} g(x, y) < t$ which is necessary to cover the support of π, then $N(\operatorname{supp}(\pi), \|\cdot - \cdot\|^p, t) = N\left(\pi, \operatorname{supp}(\pi), \|\cdot - \cdot\|, t^{1/p}\right)$. Even if π has full support on the unit ball of a D-dimensional Banach space, and g is the euclidean metric, then $N(\operatorname{supp}(\pi), \|\cdot - \cdot\|, t) = Kt^{-D}$ ([9]) and $N(\operatorname{supp}(\pi), \|\cdot - \cdot\|^p, t) = Kt^{-D/p}$, decreasing the covering number by the factor $t^{D(1-1/p)}$ which can make a big difference already for $p = 1$.

Local Lipschitz properties. Let $\mathcal{F}$ be the class of all measurable functions $f : \mathcal{X} \to [0, \infty)$, and let ρ be any extended real valued function $\rho : \mathcal{X}^2 \to [0, \infty]$ satisfying $\rho(x, y) = 0 \iff x = y$, and define $L : \mathcal{F} \times \mathcal{X} \to [0, \infty]$ by

$$L_\rho(f, x) := \sup_{y \neq x} \frac{f(y) - f(x)}{\rho(y, x)}.$$

Then with Young's inequality, for $p^{-1} + q^{-1} = 1$, and $y, x \in \mathcal{X}$ and any $f \in \mathcal{F}$

$$f(y) \leq f(x) + L_\rho(f, x)\, \rho(y, x) \leq f(x) + \frac{L_\rho(f, x)^p}{p} + \frac{\rho(y, x)^q}{q}.$$

It follows that

$$\Phi(f, x) = f(x) + \frac{L_\rho(f, x)^p}{p} \text{ and } g(y, x) = \frac{\rho(y, x)^q}{q}$$

define a gauge pair, to which Theorem 3.2 (i) can be applied and gives a high probability bound, uniform over all non-negative functions. When can we expect this bound to be small?

Suppose $(\mathcal{X}, d)$ is a metric space, let $r > 0$ and define

$$\rho(y,x) = \begin{cases} d(y,x) & \text{if} \quad d(y,x) \le r \\ +\infty & \text{if} \quad d(y,x) > r \end{cases}, \text{ so } L_\rho(f,x) = \sup_{y:0<d(y,x)\le r} \frac{f(y) - f(x)}{d(y,x)}.$$

$L_\rho(f,x)$ measures a Lipschitz-type property, localized at x with range r, which is always bounded by the local Lipschitz constant of f in the ball of radius r about x, but may be substantially smaller (take $f(x_1,x_2) = \sqrt{x_1^2 + x_2^2}\text{sign}(x_1)$ in $\mathbb{R}^2$. Then $L_\rho(f,(0,0)) = 1$ but the local Lipschitz constant is infinite on any ball of nonzero radius about $(0,0)$). With $p = q = 2$ the bound for the iid case becomes

$$\sup_{f\in\mathcal{F}} \Pr\left\{ f(X) > \max_{j\in S\cap[n-1]} f(X_j) + \frac{L_\rho(f,X_j)^2}{2} + t \mid \mathbf{X}_1^n \right\}$$
$$\le \frac{1}{n-1}\sum_{k=2}^{n} \mathbf{1}\left\{ \min_{i\in S\cap[k-1]} d(X_k,X_i)^2 > \min\left\{t, r^2\right\} \right\} + \sqrt{\frac{2\ln(1/\delta)}{n-1}}.$$

Instead of a global constraint on $\mathcal{F}$ we require for the chosen function f

1. that the empirical error $f(X_j)$ is small and the chosen function f be nearly flat ($L_\rho(f,X_j)$ small) in the neighborhoods of the sample points X_j. Wide minima are good. These are wide minima in the data-space however, in contrast to the wide minima often cited as beneficial for deep neural networks. Note that, if there aren't too many indices j where $f(X_j) + L_\rho(f,X_j)^2/2$ is large, these indices can be collected in the exception set S^c.

2. For most X_k we find some X_i with $i < k$ and $\|X_k - X_i\| \le \min\left\{\sqrt{t}, r\right\}$.

Since ρ is unbounded we can only apply Theorem 3.2 part (i) at present. For part (ii) we need a uniform bound on $\mathcal{F}$, which may simply be postulated, and a bound on ρ, which will weaken the result above. With some large $M < \infty$ define

$$\rho(y,x) := \begin{cases} d(y,x) & \text{if} \quad d(y,x) \le r \\ 2\|\mathcal{F}\|_\infty M & \text{if} \quad d(y,x) > r \end{cases}.$$

This gives $\|g\|_\infty \le 2\|\mathcal{F}\|_\infty^2 M^2$ and $L_\rho(f,X_j)$ will be replaced $\max\left\{M^{-1}, L_\rho(f,X_j)\right\}$.

7 Some open questions

- Is the worst-case estimate in Theorem 5.2 (iii) way too pessimistic in practice? The data-dependend complexity term G could be measured for different processes. This is also interesting for iid processes. Some preliminary experiments with MNIST give values of 0.23 for $g(y,x) = \|y - x\|$ and 0.06 for $g(y,x) = \|y - x\|^2$ with sample size 1000 for the character "6" and normalized euclidean distance.
- Can the bound for unconstrained classes be applied to deep neural networks? The local Lipschitz properties of various networks could be investigated at the sample points, perhaps using the methods discussed in [12] or [18].

References

[1] Alekh Agarwal and John C Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2012.

[2] Vladimir Igorevich Arnold and André Avez. *Ergodic problems of classical mechanics*, volume 9. Benjamin, 1968.

[3] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

[4] Heinz Bauer. *Probability theory*, volume 23. Walter de Gruyter, 2011.

[5] Daniel Berend and Aryeh Kontorovich. The missing mass problem. *Statistics & Probability Letters*, 82(6):1102–1110, 2012.

[6] Richard C Bradley. Basic properties of strong mixing conditions. a survey and some open questions. 2005.

[7] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[8] JR Chazottes. Hitting and returning to non-rare events in mixing dynamical systems. *Nonlinearity*, 16(3):1017, 2003.

[9] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2001.

[10] Dylan Foster, Tuhin Sarkar, and Alexander Rakhlin. Learning nonlinear dynamical systems from a single trajectory. In *Learning for Dynamics and Control*, pages 851–861. PMLR, 2020.

[11] Daniel Hsu, Aryeh Kontorovich, David A Levin, Yuval Peres, Csaba Szepesvári, and Geoffrey Wolfer. Mixing time estimation in reversible markov chains from a single sample path. 2019.

[12] Matt Jordan and Alexandros G Dimakis. Exactly computing the local lipschitz constant of relu networks. *Advances in Neural Information Processing Systems*, 33:7344–7353, 2020.

[13] Mark Kac. On the notion of recurrence in discrete stochastic processes. 1947.

[14] C. McDiarmid. Concentration. In *Probabilistic Methods of Algorithmic Discrete Mathematics*, pages 195–248, Berlin, 1998. Springer.

[15] Ron Meir. Nonparametric time series prediction through adaptive model selection. *Machine learning*, 39:5–34, 2000.

[16] Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. *Advances in Neural Information Processing Systems*, 21, 2008.

[17] Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary φ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11(2), 2010.

[18] William Piat, Jalal Fadili, Frédéric Jurie, and Sébastien da Veiga. Towards an evaluation of lipschitz constant estimation algorithms by building models with a known lipschitz constant. In *Workshop on Trustworthy Artificial Intelligence as a part of the ECML/PKDD 22 program*, 2022.

[19] Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618. PMLR, 2019.

[20] Ingo Steinwart and Andreas Christmann. Fast learning from non-iid observations. *Advances in neural information processing systems*, 22, 2009.

[21] Geoffrey Wolfer and Aryeh Kontorovich. Estimating the mixing time of ergodic markov chains. In *Conference on Learning Theory*, pages 3120–3159. PMLR, 2019.

[22] Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

A Remaining proofs

A.1 Proof of Lemma 3.3

Lemma A.1 (Lemma 3.3 re-stated)**.** *Let $R_1, ..., R_n$ be real random variables $0 \leq R_j \leq 1$ and let $\sigma_1 \subseteq \sigma_2 \subseteq ...\sigma_n$ be a filtration such that R_j is σ_j-measurable. Let $\hat{V} = \frac{1}{n}\sum_j R_j$, $V = \frac{1}{n}\sum_j \mathbb{E}\left[R_j|\sigma_{j-1}\right]$. Then*

$$\Pr\left\{V > 2\hat{V} + t\right\} \leq e^{-nt/e}$$

and equivalently, for $\delta > 0$,

$$\Pr\left\{V > 2\hat{V} + \frac{e\ln(1/\delta)}{n}\right\} \leq \delta.$$

Proof. Let $Y_j := \frac{1}{n}\left(\mathbb{E}\left[R_j|\sigma_{j-1}\right] - R_j\right)$, so $\mathbb{E}\left[Y_j|\sigma_{j-1}\right] = 0$.
Then $\mathbb{E}\left[Y_j^2|\sigma_{j-1}\right] = \left(1/n^2\right)\left(\mathbb{E}\left[R_j^2|\sigma_{j-1}\right] - \mathbb{E}\left[R_j|\sigma_{j-1}\right]^2\right) \leq (1/n)^2\,\mathbb{E}\left[R_j|\sigma_{j-1}\right]$, since $0 \leq R_j \leq 1$. Define a real function g by

$$g(t) = \frac{e^t - t - 1}{t^2} \text{ for } t \neq 0 \text{ and } g(0) = \frac{1}{2}$$

It is standard to verify that $g(t)$ is nondecreasing for $t \geq 0$ ([14]). Fix $\lambda > 0$. We have for all $x \leq \lambda$ that $e^x \leq 1 + x + g(\lambda)\,x^2$. For any β with $0 < \beta \leq n\lambda$ we then have

$$\begin{aligned}\mathbb{E}\left[e^{\beta Y_j}|\sigma_{j-1}\right] &\leq \mathbb{E}\left[1 + \beta Y_j + g(\lambda)\,\beta^2 Y_j^2|\sigma_{j-1}\right] = 1 + g(\lambda)\,\beta^2\mathbb{E}\left[Y_j^2|\sigma_{j-1}\right] \\ &\leq \exp\left(g(\lambda)\,\beta^2\mathbb{E}\left[Y_j^2|\sigma_{j-1}\right]\right) \leq \exp\left(g(\lambda)\left(\frac{\beta}{n}\right)^2\mathbb{E}\left[R_j|\sigma_{j-1}\right]\right),\end{aligned}$$

where we also used $1 + x \leq e^x$. Defining $Z_0 = 1$ and for $j \geq 1$

$$Z_j = Z_{j-1}\exp\left(\beta Y_j - g(\lambda)\left(\frac{\beta}{n}\right)^2\mathbb{E}\left[R_j|\sigma_{j-1}\right]\right)$$

then

$$\mathbb{E}\left[Z_j|\sigma_{j-1}\right] = \exp\left(-g(\lambda)\left(\frac{\beta}{n}\right)^2\mathbb{E}\left[R_j|\sigma_{j-1}\right]\right)\mathbb{E}\left[e^{\beta Y_j}|\sigma_{j-1}\right]Z_{j-1} \leq Z_{j-1}.$$

It follows that $\mathbb{E}\left[Z_n\right] \leq 1$. Spelled out this is

$$1 \geq \mathbb{E}\left[\exp\left(\beta\left(V - \hat{V}\right) - \frac{g(\lambda)\,\beta^2}{n}V\right)\right].$$

If we choose $\beta = n\lambda$ then

$$1 \geq \mathbb{E}\left[\exp\left(n\lambda\left(V - \hat{V}\right) - ng(\lambda)\,\lambda^2 V\right)\right] = \mathbb{E}\left[\exp\left(n\left(1 + 2\lambda - e^\lambda\right)V - n\lambda\hat{V}\right)\right].$$

Using calculus to maximize the coefficient $1 + 2\lambda - e^\lambda$ of V we set $\lambda = \ln 2$ and obtain

$$1 \leq \mathbb{E}\left[\exp\left(n\left(2\ln 2 - 1\right)\left(V - \frac{\ln 2}{2\ln 2 - 1}\hat{V}\right)\right)\right].$$

Markov's inequality then gives

$$\Pr\left\{V > \frac{\ln 2}{2\ln 2 - 1}\hat{V} + t\right\} \leq \exp\left(-\left(2\ln 2 - 1\right)nt\right).$$

To get the result we use $\ln 2/\left(2\ln 2 - 1\right) \leq 2$ and $2\ln 2 - 1 \geq 1/e$.

$$\Pr\left\{V > 2\hat{V} + t\right\} \leq e^{-nt/e}.$$

□

A.2 Proof of Lemma 6.1

The following is a version of Lemma 6.1, where the parameter γ is allowed to depend on x.

Lemma A.2. *Let f be a nonnegative, differentiable function on a convex open subset $\mathcal{O}$ of a Hilbert space and $\lambda > 0$. Fix $x \in \mathcal{O}$ and suppose that for $\gamma > 0$ and all $y \in \mathcal{O}$ we have*

$$\left\|f'(y) - f'(x)\right\| \leq \gamma\left\|y - x\right\|$$

Then for any $y \in \mathcal{O}$

$$f(y) \leq \left(1 + \lambda^{-1}\right)f(x) + (1 + \lambda)\,\frac{\gamma}{2}\left\|y - x\right\|^2.$$

If $f(x) = 0$ then $f(y) \leq \frac{\gamma}{2}\left\|y - x\right\|^2$.

Proof. From the fundamental theorem of calculus we get for any $x, y \in \mathcal{X}$

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \langle f'(x + t(y-x)), y - x\rangle \, dt \\ &= f(x) + \langle f'(x), y - x\rangle + \int_0^1 \langle f'(x + t(y-x)) - f'(x), y - x\rangle \, dt \\ &\leq f(x) + \langle f'(x), y - x\rangle + \frac{\gamma}{2}\|y - x\|^2. \end{aligned} \tag{2}$$

Since f is nonnegative we get for any y and fixed x

$$\langle f'(x), x - y\rangle \leq f(x) + \frac{\gamma}{2}\|y - x\|^2.$$

Letting $t > 0$ and $x - y = tf'(x) / \|f'(x)\|$ or $y = x - tf'(x) / \|f'(x)\|$ we obtain after division by t

$$\|f'(x)\| \leq \frac{f(x)}{t} + \frac{\gamma t}{2}.$$

Using calculus to minimize we find $\|f'(x)\| \leq \sqrt{2\gamma f(x)}$ and from (2) and Young's inequality

$$\begin{aligned} f(y) &\leq f(x) + 2\sqrt{\lambda^{-1} f(x) \lambda \frac{\gamma}{2}\|y - x\|^2} + \frac{\gamma}{2}\|y - x\|^2 \\ &\leq \left(1 + \lambda^{-1}\right) f(x) + (1 + \lambda)\frac{\gamma}{2}\|y - x\|^2. \end{aligned}$$

The other inequality is obtained by letting $\lambda \to 0$. □

A.3 Proof of Theorem 5.2.

For the proof we need an auxiliary construction, projecting the process onto a partion. If $\mathcal{C} = (C_1, ..., C_N)$ is any disjoint partition of $\mathcal{X}$ into measurable subsets, define a process $\mathbf{Y} = (Y_i)_{i\in\mathbb{N}}$ with values in space $[N]$ by $Y_i = j \iff X_i \in C_j$. The process $\mathbf{Y}$ inherits its ergodicity and mixing properties from $\mathbf{X}$, in the sense that $\mathbf{Y}$ is ergodic whenever $\mathbf{X}$ is, and the mixing coefficients of $\mathbf{Y}$ are bounded by the mixing coefficients of $\mathbf{X}$.

Lemma A.3. *Assume $\|g\|_\infty = 1$. Let $t \in (0,1)$, $\tau \in \mathbb{N}$, $n \geq (1 + 4/t)\tau$, and that we can cover the support of π with N disjoint measurable sets $C_1, ..., C_N$ such that $\mathrm{diam}_g(C_j) < t/2$ for all j. Then, with $m(n) = \lfloor nt/(2\tau)\rfloor$,*

$$\Pr\{G(\mathbf{X}, n, \tau, g) > t\} \leq \sum_{j=1}^{N} \Pr\{\exists k > nt/2, Y_k = j, \forall 1 \leq i \leq nt/2 - \tau, Y_i \neq j\}.$$

Proof. Write $M(n) = m(n)\tau$. Then

$$\begin{aligned} &\Pr\{G(\mathbf{X}, n, \tau, g) > t\} \\ = \;& \Pr\left\{\frac{1}{n-\tau}\sum_{k=\tau+1}^{n} \min_{i\in[k-\tau]} g(X_k, X_i) > t\right\} \\ = \;& \Pr\left\{\frac{1}{n-\tau}\sum_{k:\tau<k\leq nt/2} \min_{i:1\leq i\leq k-\tau} g(X_k, X_i) + \frac{1}{n-\tau}\sum_{k:k>nt/2}^{n} \min_{i:1\leq i\leq k-\tau} g(X_k, X_i) > t\right\} \\ \leq \;& \Pr\left\{\frac{1}{n-\tau}\sum_{k:k>nt/2}^{n} \min_{i:1\leq i\leq k-\tau} g(X_k, X_i) > \frac{t}{2}\right\} \\ \leq \;& \Pr\left\{\exists k > nt/2, \min_{i:1\leq i\leq nt/2-\tau} g(X_k, X_i) > \frac{t}{2}\right\}. \end{aligned}$$

Recall that $\text{diam}_g(C_j) < t/2$. Now assume that $X_k \in C_j$ and $\min_{i:1\leq i\leq k-\tau} g(X_k, X_i) > t/2$. Then none of the X_i can be in C_j because $\sup_{y\in C_j} g(X_k, y) \leq \text{diam}_g(C_j) < t/2$. We therefore must have $X_i \notin C_j = C(X_k)$ for all i. Thus

$$\begin{aligned}\Pr\{G(\mathbf{X}, n, \tau, g) > t\} &\leq \Pr\{\exists k > nt/2, \forall i : 1 \leq i \leq nt/2 - \tau, X_i \notin C(X_k)\} \\ &= \sum_{j=1}^{N} \Pr\{\exists k > nt/2, Y_k = j, \forall 1 \leq i \leq nt/2 - \tau, Y_i \neq j\}.\end{aligned}$$

□

Proof of Theorem 5.2. Since $\mathcal{X}$ is g-totally bounded there exists a partition $C_1, ..., C_N$ such that $\text{diam}_g(C_j) < t/2$ for all j, as required for the previous lemma. We may assume that $\pi(C_j) > 0$ and we work with the induced process $\mathbf{Y}$.

(i) Fix $\epsilon > 0$. If the underlying process is ergodic, so is the process Y_k, and for every j we have

$$\Pr\{\forall 1 \leq i \leq n, Y_i \neq j\} \to 0 \text{ as } n \to \infty.$$

Then there is $n_0 \in \mathbb{N}$ such that forall $n \geq n_0$ we have $\Pr\{\forall 1 \leq i \leq nt/2 - \tau, Y_i \neq j\} < \epsilon/N$, so by the lemma

$$\begin{aligned}\Pr\{G(\mathbf{X}, n, \tau, g) > t\} &\leq \Pr\{\exists k > nt/2, \forall i : 1 \leq i \leq nt/2 - \tau, X_i \notin C(X_k)\} \\ &\leq \Pr\{\exists j \in [N], \forall i : 1 \leq i \leq nt/2 - \tau, Y_i \neq j\} < \epsilon.\end{aligned}$$

(ii) To prove almost sure convergence we use the following consequence of the Borel-Cantelli lemma ([4]): let Z_n be a sequence of random variables. If for every $t > 0$ we have $\sum_{n>1} \Pr\{|Z_n| > t\} < \infty$ then $Z_n \to 0$ almost surely as $n \to \infty$.

Now note that the events $\{\exists k > nt/2, Y_k = j\}$ and $\{\forall i : 1 \leq i \leq nt - \tau, Y_i \neq j\}$ are separated by a time interval at least τ, so, by the definition of the α-mixing coefficients, we have from the lemma that

$$\begin{aligned}&\Pr\{G(\mathbf{X}, n, \tau, g) > t\} \\ \leq\ & \sum_{j=1}^{N} \Pr\{\exists k > nt/2, Y_k = j, \forall 1 \leq i \leq nt/2 - \tau, Y_i \neq j\} \\ \leq\ & \sum_{j} \Pr\{X \in C_j\} \Pr \bigcap_{i:1\leq i\leq \frac{nt-2\tau}{2}} \{X_i \notin C_j\} + \alpha_\tau \\ \leq\ & \sum_{j} \Pr\{X \in C_j\} \Pr \bigcap_{i:1\leq i\leq nt/(2\tau)-1} \{X_{i\tau} \notin C_j\} + \alpha_\tau \\ \leq\ & \sum_{j} \pi(C_j)(1 - \pi(C_j))^{\lfloor nt/(2\tau)\rfloor - 1} + \left\lceil \frac{nt}{2\tau} \right\rceil \alpha_\tau. \end{aligned} \tag{3}$$

Now, setting $\tau = \lceil n^p \rceil$ and $\alpha_\tau \leq A\tau^{-q}$ with $p \in (2/(1+q), 1)$, we obtain for sufficiently large n,

$$\begin{aligned}\Pr\{G(\mathbf{X}, n, \lceil n^p \rceil, g) > t\} &\leq N\left(1 - \min_j \pi(C_j)\right)^{\lfloor \frac{nt}{2\tau} \rfloor - 1} + \left\lceil \frac{nt}{2\tau} \right\rceil \alpha_\tau \\ &\leq N\left(1 - \min_j \pi(C_j)\right)^{tn^{1-p}-1} + An^{1-p-qp}t.\end{aligned}$$

Since $1 - p - qp < -1$, this expression is summable, and the conclusion follows from the Borel-Cantelli Lemma.

(iii) Theorem 1 in [5] states that for $p_1, ..., p_N \geq 0$, $\sum p_i = 1$ and $m \in \mathbb{N}$ we have

$$\sum p_i (1 - p_i)^m \leq \frac{N}{em}.$$

Applying this with (3) gives

$$\Pr\left\{G\left(\mathbf{X}, n, \tau, g\right) > t\right\} \leq \frac{N}{e\left(\lfloor nt/\left(2\tau\right)\rfloor - 1\right)} + \left\lceil \frac{nt}{2\tau} \right\rceil \alpha_{\tau}.$$

□