Generalization Bounds for Minimum Volume Set Estimation based on Markovian Data

Patrice Bertail[*], **Gabriela Ciołek**[**], **Stephan Clémençon**[***]
[*] Modal'X, Université Paris Ouest Nanterre la Défense, France
[**,***]LTCI Telecom ParisTech, Université Paris-Saclay
75013, Paris, France
[**] AGH University of Science and Technology,
al. Mickiewicza 30, 30-059 Krakow, Poland
[**] gabriela.ciolek@telecom-paristech.fr

Abstract

The main goal of this paper is to establish generalization bounds for minimum volume set estimation for regenerative Markov chains. We obtain new maximal concentration inequality in order to show that learning rate bounds depend not only on the complexity of the class of candidate sets but also on the ergodicity rate of the chain X, expressed in terms of tail conditions for the length of the regenerative cycles. Finally, we show that it is straightforward to extend the preceding results to the Harris recurrent case.

Preliminaries

Machine learning theory for dependent processes has been intensively investigated in the last years, see for instance [1], [2], [14] or [6] for the results stated in a very general setting. In statistical learning theory, numerous works established non-asymptotic bounds assessing the generalization capacity of empirical risk minimizers under a large variety of complexity assumptions for the class of decision rules over which optimization is performed, by means of sharp control of uniform deviation of i.i.d. averages from their expectation, while fully ignoring the possible dependence across training data in general. It is the purpose of this paper to show that similar results can be obtained when statistical learning is based on a data sequence drawn from a (Harris positive) Markov chain X, through the example of estimation of *minimum volume sets* (MV-sets) related to X's stationary distribution. The generalization bounds for MV-set estimation problem in the i.i.d. setting were established in [13]. Since then, the result has been extended to strong mixing processes in [4]. We aim to generalize the aforementioned bounds for a more general classes of dependent processes.

Background on Markov chain theory

Throughout the paper, $X = (X_n)_{n\in\mathbb{N}}$ is a ψ-irreducible time-homogeneous Markov chain, valued in a countably generated measurable space $(E, \mathcal{E})$ with transition probability $\Pi(x, dy)$ and initial distribution ν (refer to [12] for an account of the Markov chain theory). In addition, $\mathbb{P}_\nu$ means the probability measure on the underlying space such that $X_0 \sim \nu$. We write $\mathbb{P}_x$ when considering the Dirac mass at $x \in E$. The expectations under $\mathbb{P}_\nu$ and $\mathbb{P}_x$ are denoted by $\mathbb{E}_\nu[.]$ and $\mathbb{E}_x[.]$ respectively. We assume further that the chain X is Harris recurrent, meaning that the chain visits an infinite number of times any subset $B \in \mathcal{E}$ such that $\psi(B) > 0$ with probability one whatever the initial state, ψ being a maximal irreducibility measure, *i.e.* $\mathbb{P}_x(\sum_{n\geq 1} \mathbb{I}\{X_n \in B\} = \infty) = 1$, for all $x \in E$. Within this framework, a Markov chain is said to be *regenerative* when it possesses an accessible atom

Definition 1. *Assume that X is aperiodic and ψ-irreducible. We say that a set $A \in \mathcal{E}$ is an accessible atom if for all $x,\ y \in A$ we have $\Pi(x, \cdot) = \Pi(y, \cdot)$ and $\psi(A) > 0$.*

We say that X is positive recurrent if and only if the expected return time to the atom is finite, *i.e.* $\mathbb{E}_A[\tau_A] < \infty$. Then, it follows from the Kac's theorem that the invariant probability distribution μ is the Pitman's *occupation measure* given by

$$\mu(B) = \frac{1}{\mathbb{E}_A[\tau_A]}\mathbb{E}_A\left[\sum_{i=1}^{\tau_A}\mathbb{I}\{X_i \in B\}\right], \text{ for all } B \in \mathcal{E}. \tag{1}$$

We define the sequence of regeneration times $(\tau_A(j))_{j\geq 1}$. Let $\tau_A = \tau_A(1) = \inf\{n \geq 1 : X_n \in A\}$ and $\tau_A(j) = \inf\{n > \tau_A(j-1), X_n \in A\}$ for $j \geq 2$. By the strong Markov property, given any initial law ν, the sample paths of X can be divided into i.i.d. blocks corresponding to consecutive visits of the chain to atom A. The segments of data are of the form: $\mathcal{B}_j = (X_{1+\tau_A(j)}, \cdots, X_{\tau_A(j+1)}),\ j \geq 1$ and take values in the torus $\cup_{k=1}^{\infty} E^k$.

We introduce few more pieces of notation: we write $l_n = \sum_{i=1}^{n} \mathbb{I}\{X_i \in A\}$ for the total number of consecutive visits of the chain to the atom A. We make the convention that $B_{l_n}^{(n)} = \emptyset$ when $\tau_A(l_n) = n$. Denote by $l(B_j) = \tau_A(j+1) - \tau_A(j),\ j \geq 1$, the length of regeneration blocks.

In this framework we also consider more general class of Markov chains, namely Harris recurrent Markov chains.

Definition 2. *Assume that X is a ψ-irreducible Markov chain. We say that X is Harris recurrent iff, starting from any point $x \in E$ and any set such that $\psi(A) > 0$, we have*

$$\mathbb{P}_x(\tau_A < +\infty) = 1.$$

Observe that the property of Harris recurrence ensures that X visits set A infinitely often a.s. (see also [7] for

more details). Suppose that X is a positive recurrent Harris Markov chain. By the results of Nummelin in [8], it is possible to generalize the theory from atomic case to general Harris case through the construction of an artificial atom. The *Nummelin splitting technique* relies on the existence of so-called small sets, which have the following property: there exists a parameter $\delta > 0$, a positive probability measure Φ supported by S and an integer $m \in N^*$ such that

$$\forall x \in S,\ A \in \mathcal{E}\ \ \Pi^m(x, A) \geq \delta\, \Phi(A), \tag{2}$$

where Π^m designates the m-th iterate of the transition probability Π. We call (2) the minorization condition and denote by $\mathcal{M}$.

The Nummelin technique. We now explain how to construct the atomic chain onto which the initial chain X is embedded. Suppose that X satisfies $\mathcal{M} = \mathcal{M}(m, S, \delta, \Psi)$ for $S \in \mathcal{E}$ such that $\psi(S) > 0$. Rather than replacing the initial chain X by the chain $\{(X_{nm}, ..., X_{n(m+1)-1})\}_{n\in\mathbb{N}}$, we suppose $m = 1$. The sample space is expanded so as to define a sequence $(Y_n)_{n\in\mathbb{N}}$ of independent Bernoulli r.v.'s with parameter δ by defining the joint distribution $\mathbb{P}_{\nu,\mathcal{M}}$ whose construction relies on the following randomization of the transition probability Π each time the chain hits S. Note that it occurs with probability one since the chain is Harris recurrent and $\psi(S) > 0$. If $X_n \in S$, and

- if $Y_n = 1$ (occurs with probability $\delta \in\ ]0, 1[$), then $X_{n+1} \sim \Phi$,
- if $Y_n = 0$, then $X_{n+1} \sim (1-\delta)^{-1}(\Pi(X_n, .) - \delta\Phi(.))$.

Set $\text{Ber}_\delta(\beta) = \delta\beta + (1-\delta)(1-\beta)$ for $\beta \in \{0, 1\}$. We have thus constructed the *split chain* $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$, valued in $E \times \{0, 1\}$ with transition kernel $\Pi_{\mathcal{M}}$ defined by

- for any $x \notin S$, $B \in \mathcal{E}$, β and β' in $\{0, 1\}$,

$$\Pi_{\mathcal{M}}\left((x, \beta), B \times \{\beta'\}\right) = \Pi\left(x, B\right) \times \text{Ber}_\delta(\beta'),$$

- for any $x \in S$, $B \in \mathcal{E}$, β' in $\{0, 1\}$,
  - $\Pi_{\mathcal{M}}\left((x, 1), B \times \{\beta'\}\right) = \Phi(B) \times \text{Ber}_\delta(\beta')$
  - $\Pi_{\mathcal{M}}\left((x, 0), B \times \{\beta'\}\right) = (1-\delta)^{-1}(\Pi\left(x, B\right) - \delta\Phi(B)) \times \text{Ber}_\delta(\beta')$.

The key point of the construction relies on the fact that $A_S = S \times \{1\}$ is the atom for the bivariate chain (X, Y) which inherits all communication and stochastic stability properties from X.

Minimum Volume Set Estimation

The notion of Minimum Volume set (MV-set) has been proposed in [5] in order to extend the definition of quantile for 1-dimensional probability distributions. Consider a probability distribution μ on a measurable space $(E, \mathcal{E})$. Let $\alpha \in (0, 1)$ and λ be a σ-finite measure of reference on $(E, \mathcal{E})$, any solution of the minimization problem

$$\min_{\Omega\in\mathcal{E}} \lambda(\Omega) \text{ subject to } \mu(\Omega) \geq \alpha \tag{3}$$

is called a MV-set of level α. Throughout the paper, we assume that the distribution μ is absolutely continuous w.r.t. λ and denote by $f(x) = (d\mu/d\lambda)(x)$ the related density. For any $\alpha \in (0, 1)$, under the assumptions that the density f is bounded and that the image of μ by f, denoted by μ_f, is a continuous probability on $\mathbb{R}_+$, it is shown in [11] that the set $\Omega_\alpha^* = \{x \in E :\ f(x) \geq \mu_f^{-1}(1-\alpha)\}$ is the unique solution of the MV-set estimation problem (3). For small values of the mass level α, MV-sets should permit to recover the modes of the distribution μ, while their complementary sets correspond to *rare observations* when α is large.

Empirical MV-sets in the i.i.d. setting. A level $\alpha \in (0, 1)$ being preliminarily fixed, a natural way of building estimates of the MV set Ω_α^* from the i.i.d. data $X_1, \ldots, X_n \sim \mu(dx)$ consists in solving a statistical version of the constrained optimization problem (3)

$$\min_{\Omega\in\mathcal{G}} \lambda(\Omega) \text{ subject to } \widehat{\mu}_n(\Omega) \geq \alpha - \psi_n, \tag{4}$$

where the empirical distribution $\widehat{\mu}_n = (1/n)\sum_{i=1}^n \delta_{X_i}$ (or a smoothed counterpart of the latter) replaces the unknown probability measure μ, minimization is restricted to a subset $\mathcal{G}$ of $\mathcal{E}$, expected to be sufficiently rich to include a reasonable approximation of Ω_α^*, and ψ_n plays the role of a *tolerance parameter*, chosen of the same order of magnitude as the supremum $\sup_{\Omega\in\mathcal{G}} |\widehat{\mu}_n(\Omega) - \mu(\Omega)|$. This approach, that essentially boils down to substituting the true (unknown) probability measure $\mu(dx)$ with its statistical counterpart is referred to as MV-ERM in [13]. The class $\mathcal{G}$ is ideally made of sets $\Omega \in \mathcal{E}$ whose *volume* $\lambda(\Omega)$ can be efficiently computed or estimated, *e.g.* by Monte-Carlo simulation.

Main Results

We now state the main results of the paper, related to the performance of solutions of the problem (4) when the empirical probability estimates $\widehat{\mu}_n(\Omega)$ are based on a Markovian trajectory of length $n \geq 1$. For simplicity, we assume that $E \subset \mathbb{R}^d$ with $d \geq 1$ and that $\lambda(dx)$ is the restriction of Lebesgue measure on E equipped with its Borel σ-algebra. We first address the case where the underlying Markov chain is regenerative and explain next how this apparently restrictive result can be straightforwardly extended to general positive recurrent chains.

We start with considering the situation where the positive recurrent Markov chain X possesses an accessible atom A. Its stationary distribution is then given by (1) and its empirical counterpart based on the sequence $X_1, \ldots, X_n$ can be rewritten as: $\forall\ \Omega \in \mathcal{E}$,

$$\widehat{\mu}_n(\Omega) = \frac{1}{n}\sum_{i=1}^{\tau_A} \mathbb{I}\{X_i \in \Omega\} + \frac{l_n - 1}{n}\left(\frac{1}{l_n - 1}\sum_{j=1}^{l_n-1} S_j(\Omega)\right) + \frac{1}{n}\sum_{i=1+\tau_A(l_n)}^{n} \mathbb{I}\{X_i \in \Omega\}, \tag{5}$$

where the occupation time of the set Ω between the j-th and $(j+1)$-th regeneration times is denoted by $S_j(\Omega) = \sum_{\tau_A(j)<i\leq\tau_A(j+1)} \mathbb{I}\{X_i \in \Omega\}$ for $j \geq 1$ and with the usual convention that empty summation is equal to zero. Under the assumptions we made, the $S_j(\Omega)$'s are integrable i.i.d.

r.v.'s with common mean $\mathbb{E}_A[\tau_A]\mu(A)$, $l_n \sim n/\mathbb{E}_A[\tau_A]$ almost-surely as $n \to +\infty$ and the first and last terms in the equation above both almost-surely asymptotically vanish. Hence, the random variables $S_j(\Omega)/\mathbb{E}_A[\tau_A]$ shall play the role of training observations in the subsequent analysis: the smaller the expected cycle length $\mathbb{E}[\tau_A]$, the larger the probability to observe a high number of training observations. However, one must pay attention to the fact that the $l_n - 1$ regeneration data blocks, though asymptotically i.i.d., are not independent. Except for the i.i.d. situation (notice that in such case, the whole state space E can be viewed as an atom), the frequency of visits to a candidate set Ω over the path $X_1, \ldots, X_n$ is not an i.i.d. average. Decomposition (5) is the main ingredient to control $\sup_{\Omega\in\mathcal{G}} |\widehat{\mu}_n(\Omega) - \mu(\Omega)|$. We will need the following assumptions in the subsequent analysis. Let $p \geq 2$.

Assumption 1. *The collection of indicator functions on E, $\mathcal{F} = \{\mathbb{I}\{x \in \Omega\} : \Omega \in \mathcal{G}\}$ is a uniform Donsker class (relative to L_1) with polynomial uniform covering numbers*[1]*, i.e. there exists a constant $c > 0$ s.t. $\forall \zeta > 0$,*

$$\mathcal{N}_1(\zeta, \mathcal{F}) \stackrel{def}{=} \sup_Q \mathcal{N}(\zeta, \mathcal{F}, L_1(Q)) \leq c(1/\zeta)^p,$$

where the supremum is taken over the set of finitely discrete probability measures on $(E, \mathcal{E})$.

Assumption 2. *We have: $\mathbb{E}_A[\tau_A^p] < \infty$, $\mathbb{E}_\nu[\tau_A^p] < \infty$ and $\mathbb{E}_A[l(B_1)]^p < \infty$.*

Under Assumption 2, the ergodicity rate of the chain X is at least subgeometric, polynomial namely, in the sense that $\sup_{h:\, ||h||_\infty \leq 1} |h(X_n) - \mu(h)| = O(1/n^{p-1})$, see [16]. As the following theorem shows, Assumption 2 combined with Assumption 1 allows to control the fluctuations of (5) uniformly over $\mathcal{G}$. In addition, we point out that the degree $p \geq 2$ involved in Assumptions 1 and 2 is the same here, for the sole purpose of establishing learning rate for the MV-ERM method.

Before we establish generalization bounds for the MV-set, we state a new concentration inequality since the proof of our generalization bounds heavily relies on it. We indicate that our inequality may be used to show generalization bounds of numerous machine learning algorithms for Markovian data under appropriate change of the underlying assumptions such as block moment assumptions or a complexity condition imposed on the considered class of functions $\mathcal{F}$. Let $\sigma_m^2 = \max_{\Omega\in\mathcal{G}} \sigma^2(S_j(\Omega))$.

Lemma 1. *Let $\zeta > 0$ and $p \geq 2$. Suppose that Assumptions 1-2 are fulfilled and $\sup_Q \mathcal{N}(\zeta/120, \mathcal{F}, L_1(Q)) < +\infty$. Then, for all $\zeta > 0$, we have: $\forall n \geq \frac{60^2\sigma_m^2}{\zeta^2}$ and for any $M_l > 0$*

$$\mathbb{P}_\nu\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{l_n} S_i(\Omega) - \mathbb{E}_\mu(S_i(\Omega))\right| \geq \zeta\right\} \leq \frac{\mathcal{N}_1(\zeta_1, \mathcal{F}) \times C_1}{\zeta^p \times n^{p/2-1}} + C_2\mathbb{P}_A\left(\sum_{i=1}^{n} l(B_i) > U_n\right),$$

where $\zeta_1 = \frac{\zeta}{120M_l\mathbb{E}_A[\tau_A]}$, $U_n = nM_l\mathbb{E}_A[l(B_1])$ and C_1, C_2 are constants which can be explicitly computed and are given in the proof.

Remark 1. *Note that it is easy to obtain a bound for the probability $\mathbb{P}_A\left(\sum_{i=1}^n l(B_i) > U_n\right)$ ($U_n = nM_l\mathbb{E}_A[l(B_1)]$) since $\sum_{i=1}^n l(B_i)$ is a sum of i.i.d. random variables. Under Assumption 2 it follows directly from Theorem 2.10 in [9] that for any $M_l > 1$*

$$\mathbb{P}_A\left(\sum_{i=1}^{n} l(B_i) > U_n\right) \leq \frac{K}{n^{p/2}(M_l - 1)^p},$$

where $K = \frac{C\mathbb{E}_A[l(B_1)]^p}{(\mathbb{E}_A[l(B_1)])^p}$ and C is a positive constant that depends only on p. Obviously, under sharper moment conditions imposed on $l(B_i)'s$, one can obtain exponential bounds for the aforementioned probability.

Observe that the larger the degree p that controls in particular the decay of the tail of the distribution of the regeneration cycle length (and X's ergodicity rate) and thus the distribution of the number of cycles within a trajectory of finite length, the smaller the rate bound for the maximal deviation (and thus the learning rate of the MV-ERM method, see the result below). For $p \to +\infty$, one asymptotically recovers the usual i.i.d. bound.

Proof of Lemma 1. Firstly, we deal with the random number of blocks l_n which is correlated with the blocks itself. By the Montgomery-Smith inequality (see [3], Theorem 1.1.5) we get immediately that

$$\mathbb{P}_\nu\left\{\sup_{\Omega\in\mathcal{G}}\left|\sum_{j=1}^{l_n-1}\{S_j(\Omega) - \mathbb{E}_\mu[S_j(\Omega)]\}\right| \geq \zeta\right\} \leq 9\mathbb{P}_A\left\{\sup_{\Omega\in\mathcal{G}}\left|\sum_{j=1}^{n}\{S_j(\Omega) - \mathbb{E}_\mu[S_j(\Omega)]\}\right| \geq \zeta/30\right\}. \tag{6}$$

We now explain how to apply standard arguments in empirical processes theory to the latter.

Ghost sample of regeneration blocks and symmetrization. Consider an i.i.d. sample $(\mathcal{B}_1', \ldots, \mathcal{B}_n')$, independent copy of the sample of regenerative blocks $(\mathcal{B}_1, \ldots, \mathcal{B}_n)$. The corresponding sample of block variables

$$S' = (S_1'(\Omega), \ldots, S_n'(\Omega))$$

is an independent copy of

$$S = (S_1(\Omega), \ldots, S_n(\Omega)) \quad \textit{for any } \Omega \in \mathcal{G}.$$

[1] Recall that, for any $\zeta > 0$ and probability measure Q, the covering number $\mathcal{N}(\zeta, \mathcal{F}, L_1(Q))$ is the minimal number of $L_1(Q)$ balls of radius ζ needed to cover the class $\mathcal{F}$.

Using the symmetrization lemma in [10] (see p. 14 therein), we have: $\forall \zeta > 0$,

$$\mathbb{P}_A \left\{ \sup_{\Omega \in \mathcal{G}} \left| \frac{1}{n} \sum_{j=1}^{n} (S_j(\Omega) - \mathbb{E}_\mu(S_j(\Omega))) \right| \geq \frac{\zeta}{30} \right\} \leq \frac{1}{\beta} \mathbb{P}_A \left(\sup_{\Omega \in \mathcal{G}} \left| \frac{1}{n} \sum_{j=1}^{n} (S_j(\Omega) - S'_j(\Omega)) \right| \geq \frac{\zeta}{60} \right),$$

where $\beta = 1 - 60^2 \sigma_m^2/(n\zeta^2)$. In order to ensure that $\beta > 0$, assume that $n > 60^2 \sigma_m^2/\zeta^2$.

Randomization. Let $\epsilon_1, \ \ldots, \ \epsilon_n$ be independent Rademacher variables, independent from the $(\mathcal{B}_j, \mathcal{B}'_j)$'s. We clearly have: $\forall \zeta > 0$, $\forall n > 60^2 \sigma_m^2/\zeta^2$ and for any $M_l > 0$,

$$\begin{aligned}
&\mathbb{P}_A \left\{ \sup_{\Omega \in \mathcal{G}} \left| \frac{1}{n} \sum_{j=1}^{n} \{S_j(\Omega) - S'_j(\Omega)\} \right| \geq \frac{\zeta}{60} \right\} \\
&= \mathbb{P}_A \left\{ \sup_{\Omega \in \mathcal{G}} \left| \frac{1}{n} \sum_{j=1}^{n} \left(S_j(\Omega) - S'_j(\Omega)\right) \epsilon_i \right| \geq \frac{\zeta}{60} \right\} \\
&\leq 2\mathbb{P}_A \left(\sup_{\Omega \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} S_j(\Omega) \epsilon_i \right| \geq \frac{\zeta}{60}, \sum_{i=1}^{n} l(B_i) \leq U_n \right) \\
&\quad + 2\mathbb{P}_A \left(\sum_{i=1}^{n} l(B_i) > U_n \right),
\end{aligned}$$

where $U_n = nM_l \mathbb{E}_A l(B_i)$. In what follows, we will concentrate on the analysis of the first term in the right hand side of the inequality above.

Classes of functions on the torus. Observe that one may naturally assign to any measurable function $f : E \to \mathbb{R}$ a function $f_{\mathbb{T}}$ on the torus $\mathbb{T} = \cup_{n \geq 1} E^n$, defined by

$$f_{\mathbb{T}}(b) = f_{\mathbb{T}}(x_1, \ \ldots, \ x_n) = \sum_{i=1}^{n} f(x_i)$$

for any $n \geq 1$, $b = (x_1, \ \ldots, \ x_n) \in E^n$, which is measurable when $\mathbb{T}$ is equipped with the σ-field generated by the finite cartesian products of borelian subsets of E. Thus, we may associate the class of indicator functions

$$\mathcal{F} = \{\mathbb{I}\{x \in \Omega\} : \ \Omega \in \mathcal{G}\} \ \textit{on} \ E$$

with

$$\mathcal{F}_{\mathbb{T}} = \{f_{\mathbb{T}} : \ f \in \mathcal{F}\} = \left\{ b \in \mathbb{T} \mapsto \sum_{i=1}^{l(b)} \mathbb{I}\{b_i \in \Omega\}, \ \Omega \in \mathcal{G} \right\}.$$

Note that over the set of blocks such that

$$\frac{1}{n} \sum_{i=1}^{n} l(B_i) \leq M_l \mathbb{E}_A l(B_i)$$

we have the following $\forall (f, g) \in \mathcal{F}^2$

$$\begin{aligned}
\|f_{\mathbb{T}-g_{\mathbb{T}}}\|_{L_n^{\mathbb{T}}} &= \frac{1}{n} \sum_{i=1}^{n} |f(b_i) - g(b_i)| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \sum_{j=\tau_A(i)}^{\tau_A(i+1)} |f(x_j) - g(x_j)| \\
&\leq M_l \mathbb{E}_A l(B_i) \frac{1}{\sum_{i=1}^{n} l(B_i)} \sum_{j=1}^{\sum_{i=1}^{n} l(B_i)} |f(x_j) - g(x_j)| \\
&= M_l \mathbb{E}_A[\tau_A] \|f - g\|_{L_{1,\hat{\mathbb{P}}_n}}, \qquad (7)
\end{aligned}$$

where

$$\hat{\mathbb{P}}_n = \frac{1}{\sum_{i=1}^{n} l(b_i)} \sum_{j=1}^{\sum_{i=1}^{n} l(b_i)} \delta_{x_j}.$$

It follows from (7) that for any $\zeta > 0$

$$\begin{aligned}
\mathcal{N}_1\left(\zeta, \mathcal{F}_{\mathbb{T}}, \mathbb{P}_n^{\mathbb{T}}\right) &\leq \mathcal{N}_1\left(\frac{\zeta}{M_l \mathbb{E}_A[\tau_A]}, \mathcal{F}, \hat{\mathbb{P}}_n\right) \\
&\leq \mathcal{N}_1\left(\frac{\zeta}{M_l \mathbb{E}_A[\tau_A]}, \mathcal{F}\right).
\end{aligned}$$

Let $h_1, h_2, \cdots, h_N$ be a collection of measurable functions on $\mathbb{T}$ such that any $f_{\mathbb{T}} \in \mathcal{F}_{\mathbb{T}}$ is at distance less than $\zeta/120$ w.r.t. the L_1-norm related to the empirical measure of the blocks $\mathcal{B}_1, \ \ldots, \ \mathcal{B}_n$ and $N \leq \mathcal{N}_1(\zeta/120, \mathcal{F})$. We have:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{B}} \left[\mathbb{P}_\epsilon \left\{ \sup_{\Omega \in \mathcal{G}} \left| \frac{1}{n} \sum_{j=1}^{n} (S_j(\Omega)) \epsilon_i \right| \geq \frac{\zeta}{60}, \sum_{i=1}^{n} l(B_i) \leq U_n \right\} \right] \\
&\leq \mathbb{E}_{\mathcal{B}} \left[\mathbb{P}_\epsilon \left\{ \max_j \left| \frac{1}{n} \sum_{i=1}^{n} (h_j(\mathcal{B}_i)) \epsilon_i \right| \geq \frac{\zeta}{120}, \sum_{i=1}^{n} l(B_i) \leq U_n \right\} \right] \qquad (8) \\
&\leq \mathcal{N}_1(\zeta_1, \mathcal{F}) \max_j \mathbb{E}_{\mathcal{B}} \mathbb{P}_\epsilon \left\{ \max_j \left| \frac{1}{n} \sum_{i=1}^{n} (h_j(\mathcal{B}_i)) \epsilon_i \right| \geq \frac{\zeta}{120} \right\}, \qquad (9)
\end{aligned}$$

where $\zeta_1 = \frac{\zeta}{120 M_l \mathbb{E}_A[\tau_A]}$ and $1 \leq j \leq \mathcal{N}_1(\frac{\zeta}{120}, \mathbb{G}_{\mathbb{T}}, \mathbb{P}_{n,\mathbb{T}})$. Finally, we apply to (8) the inequality given in Theorem 2.10 in [9] and get the following bound:

$$\begin{aligned}
&\mathbb{P}_\nu \left\{ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{l_n} S_i(\Omega) - \mathbb{E}_\mu(S_i(\Omega)) \right| \geq \zeta \right\} \\
&\leq \mathcal{N}_1\left(\zeta_1, \mathcal{F}\right) \times \frac{36 \times 5760^p e}{\beta \zeta^p \times n^{\frac{p}{2}-1}} \mathbb{E}_A[\tau_A]^p
\end{aligned}$$

that yields the result. We take $C_1 = \frac{36 \times 5760^p e \mathbb{E}_A[\tau_A^p]}{\beta}$ and $C_2 = \frac{18}{\beta}$. □

Now we are ready to state our main theorem.

Theorem 1. *Let $p \geq 2$. Suppose that Assumptions 1-2 are fulfilled. For all $\delta = \delta_1 + \delta_2 + \delta_3 + \delta_4 \in (0, 1)$, we have*

with probability at least $1-\delta:\ \forall\zeta>0\ \ \forall n\geq\frac{2\times 360^2\sigma_m^2}{\zeta^2}$,

$$\begin{aligned}&\sup_{\Omega\in\mathcal{G}}|\widehat{\mu}_n(\Omega)-\mu(\Omega)|\\&\leq\max\left[\left(\frac{K_1}{\delta_1 n^p}\right)^{1/p},\left(\frac{K_2}{\delta_2 n^p}\right)^{1/p},\right.\\&\left.\left(\frac{K_3}{n^{-1}(n^{p/2}\delta_3-2^pK_3}\right)^{1/2p},\left(\frac{K_4}{\delta_4 n^{p/2-1}}\right)^{1/2p}\right],\end{aligned}\tag{10}$$

where K_1,K_2,K_3,K_4 *are constants depending on* p, $\mathbb{E}_A[\tau_A^p]$, $\mathbb{E}_\nu[\tau_A^p]$, $\mathbb{E}_A[l(B_1)]^p$ *and the complexity and are specified in the proof.*

Proof of Theorem 1. Firstly, we do the following decomposition:

$$\begin{aligned}&\mathbb{P}_\nu\left\{\sup_{\Omega\in\mathcal{G}}|\widehat{\mu}_n(\Omega)-\mu(\Omega)|\geq\zeta\right\}\\&\leq\mathbb{P}_\nu\left\{\sup_{\Omega\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{\tau_A}\mathbb{I}\{X_i\in\Omega\}-\mu(\Omega)\right|\geq\frac{\zeta}{3}\right\}\\&+\mathbb{P}_A\left\{\sup_{\Omega\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{l_n-1}\{S_j(\Omega)-\mathbb{E}_\mu[S_j(\Omega)]\}\right|\geq\frac{\zeta}{3}\right\}\\&+\mathbb{P}_A\left\{\sup_{\Omega\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=\tau_A(l_n)+1}^{\tau_A(l_n+1)}\mathbb{I}\{X_i\in\Omega\}-\mu(\Omega)\right|\geq\frac{\zeta}{3}\right\}.\end{aligned}\tag{11}$$

The first and the last terms on the right hand side of the inequality can be easily controlled by the Chebyshev's inequality, i.e.

$$\mathbb{P}_\nu\left\{\sup_{\Omega\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{\tau_A}\mathbb{I}\{X_i\in\Omega\}-\mu(\Omega)\right|\geq\frac{\zeta}{3}\right\}\leq\frac{K_1}{n^p\zeta^p},$$

where $K_1=6^p\mathbb{E}_\nu[\tau_A^p]$ and

$$\mathbb{P}_A\left\{\sup_{\Omega\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=\tau_A(l_n)+1}^{n}\mathbb{I}\{X_i\in\Omega\}-\mu(\Omega)\right|\geq\frac{\zeta}{3}\right\}\leq\frac{K_2}{n^p\zeta^p},$$

observing that the length of the last block is less than $\tau_A(l_n+1)-\tau_A(l_n)$ and complementing the data up to the next regeneration time $\tau_A(l_n+1)+1$. We take $K_2=6^p\mathbb{E}_A[\tau_A^p]$. The control of the middle term in (11) follows directly from Lemma 1. We combine Assumption 1 with Remark 1 and obtain for any $M_l>1$

$$\begin{aligned}&\mathbb{P}_\nu\left\{\sup_{\Omega\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{l_n-1}(S_i(\Omega)-\mathbb{E}_\mu(S_i(\Omega)))\right|\geq\zeta/3\right\}\\&\leq\frac{C_1M_l^p}{\zeta^{2p}\times n^{p/2-1}}+\frac{C_2}{n^{p/2}(M_l-1)^p}\end{aligned}\tag{12}$$

where $\zeta_1=\frac{\zeta}{360M_l\mathbb{E}_A[\tau_A]}$ $C_1=[36c\times 360^p\times 3^p\times 5760^p e\mathbb{E}_A[\tau_A]^{2p}]/[1-360^2\sigma_m^2/(n\zeta^2)]$, $C_2=(18\times 72^p e\mathbb{E}_A[l(B_i)]^p)/[(1-360^2\sigma_m^2/(n\zeta^2))\times(\mathbb{E}_A l(B_1))^p]$. We note that $1/[1-360^2\sigma_m^2/(n\zeta^2)]\leq 2$ for $n\geq 2\times 360^2\sigma_m^2\zeta^2$. For simplicity's sake we consider the case when $n\geq 2\times 360^2\sigma_m^2\zeta^2$. Next, we optimize (12) in terms of M_l and get

$$M_l=1+\zeta^2 n^{-1/p}.\tag{13}$$

We solve the following equations for ζ with M_l given by (13):

$$\begin{aligned}\delta_1&=\frac{K_1}{\zeta\times n^p},\\\delta_2&=\frac{K_2}{\zeta\times n^p}\\\delta_3&=\frac{K_3M_l^p}{\zeta^{2p}\times n^{p/2-1}},\\\delta_4&=\frac{K_4}{n^{p/2}(M_l-1)^p},\end{aligned}$$

where $K_3=72\times 6^p\times 360^p\times c\times 5760^p e\mathbb{E}_A[\tau_A]^{2p}$, $K_4=18\times 72^p e\mathbb{E}_A[l(B_1)]^p/(\mathbb{E}_A l(B_1))^p$. Easy calculations show that the desired bound can be established by taking for $n\geq 2\times 360^2\sigma_m^2\zeta^2$:

$$\begin{aligned}\zeta\leq\max&\left[\left(\frac{K_1}{\delta_1 n^p}\right)^{1/p},\left(\frac{K_2}{\delta_2 n^p}\right)^{1/p},\right.\\&\left.\left(\frac{K_3}{n^{-1}(n^{p/2}\delta_3-2^pK_3}\right)^{1/2p},\left(\frac{K_4}{\delta_4 n^{p/2-1}}\right)^{1/2p}\right].\end{aligned}$$

□

A direct application of Theorem 1 to the MV-set estimation problem yields the following result.

Theorem 2. *Suppose that assumptions of Theorem 1 are fulfilled. Then, for all* $\delta=\delta_1+\delta_2+\delta_3+\delta_4\in(0,1)$, *any solution* $\widehat{\Omega}_n$ *of* (4) *with*

$$\begin{aligned}\psi_n(\delta)\stackrel{def}{=}\max&\left[\left(\frac{K_1}{\delta_1 n^p}\right)^{1/p},\left(\frac{K_2}{\delta_2 n^p}\right)^{1/p},\right.\\&\left.\left(\frac{K_3}{n^{-1}(n^{p/2}\delta_3-2^pK_3}\right)^{1/2p},\left(\frac{K_4}{\delta_4 n^{p/2-1}}\right)^{1/2p}\right]\end{aligned}$$

satisfies, with probability at least $1-\delta$,

$$\lambda(\widehat{\Omega}_n)\leq\lambda(\Omega_\alpha^*)+\left\{\inf_{\Omega\in\mathcal{G}:\,\mu(\Omega)\geq\alpha}\lambda(\Omega)-\lambda(\Omega_\alpha^*)\right\}$$

and

$$\mu(\widehat{\Omega}_n)\geq\alpha-2\psi_n(\delta).$$

Constants K_1,K_2,K_3 *and* K_4 *are given in the proof of Theorem 1.*

Proof. The proof is analogous to that of Theorem 11 in [13]. Define

$$\begin{aligned}\Theta_\mu&=\{S:\mu(\widehat{\Omega}_n)<\alpha-2\psi(S,\delta)\},\\\Gamma_\mu&=\{S:\sup_{\Omega\in\mathcal{G}}|\widehat{\mu}_n(\Omega)-\mu(\Omega)|-\psi(S,\delta)>0\},\\\mathcal{G}_\alpha&=\{\Omega\in\mathcal{G}:\mu(\Omega)\geq\alpha\}.\end{aligned}$$

Note that $S \in \overline{\Gamma}_\mu$ implies $\alpha - \mu(\hat{\Omega}_n) \leq 2\psi(S,\delta)$. Next, observe that when $S \in \Gamma_\mu^C$, with probability at least $1-\delta$ we have:

$$\lambda(\widehat{\Omega}_n) \leq \lambda(\Omega_\alpha^*) + \left\{ \inf_{\Omega \in \mathcal{G}:\, \mu(\Omega) \geq \alpha} \lambda(\Omega) - \lambda(\Omega_\alpha^*) \right\}$$

which yields the proof. □

It is straightforward to extend the preceding results into a Harris recurrent case when the regeneration properties for Harris chains can be recovered via the Nummelin splitting technique.

Assumption 3. *We have:* $\sup_{x \in S} \mathbb{E}_x[\tau_S^p] < \infty$, $\mathbb{E}_\nu[\tau_S^p] < \infty$ *and* $\sup_{x \in S} \mathbb{E}_x[l(B_1)]^p < \infty$.

It is noteworthy that the hypothesis above is independent from the small set chosen. In addition, this assumption implies that (X, Y) automatically fulfills Assumption 2, refer to Chapter 14 in [7] for further details. Recall also that these conditions can be replaced by Foster-Lyapunov drift conditions that are much more tractable in practice, see *e.g.* Chapter 11 in [7]. The following theorem gives a generalization bound for MV-set estimation problem in a Harris recurrent case. The proof is analogous to the proof of Theorem 2 (with obvious modifications as we apply the same arguments to the split chain).

Theorem 3. *Suppose that Assumptions 1-3 are fulfilled. Then, for all* $\delta = \delta_1 + \delta_2 + \delta_3 + \delta_4 \in (0,1)$*, any solution* $\widehat{\Omega}_n$ *of* (4) *with*

$$\psi_n(\delta) \stackrel{def}{=} \max \left[\left(\frac{K_1}{\delta_1 n^p} \right)^{1/p}, \left(\frac{K_2}{\delta_2 n^p} \right)^{1/p}, \left(\frac{K_3}{n^{-1}(n^{p/2}\delta_3 - 2^p K_3} \right)^{1/2p}, \left(\frac{K_4}{\delta_4 n^{p/2-1}} \right)^{1/2p} \right]$$

satisfies, with probability at least $1-\delta$*,*

$$\lambda(\widehat{\Omega}_n) \leq \lambda(\Omega_\alpha^*) + \left\{ \inf_{\Omega \in \mathcal{G}:\, \mu(\Omega) \geq \alpha} \lambda(\Omega) - \lambda(\Omega_\alpha^*) \right\}$$

and

$$\mu(\widehat{\Omega}_n) \geq \alpha - 2\psi_n(\delta).$$

The constants K_1, K_2, K_3 *and* K_4 *can be explicitly computed and are given in the proof.*

Proof. The proof boils down to solving the set of equations for ζ :

$$\delta_1 = \frac{K_1}{\zeta \times n^p},$$
$$\delta_2 = \frac{K_2}{\zeta \times n^p}$$
$$\delta_3 = \frac{K_3 M_l^P}{\zeta^{2p} \times n^{p/2-1}},$$
$$\delta_4 = \frac{K_4}{n^{p/2}(M_l - 1)^p},$$

for $n \geq 2 \times 360^2 \sigma_m^2 \zeta^2$ and with $K_1 = 6^p \mathbb{E}_\nu[\tau_S^p]$, $K_2 = 6^p \sup_{x \in S} \mathbb{E}_x[\tau_S^p]$, $K_3 = 72 \times 6^p \times 360^p \times c \times 5760^p e \sup_{x \in S} \mathbb{E}_x[\tau_S]^{2p}$ and $K_4 = 18 \times 72^p e \sup_{x \in S} \mathbb{E}_x[l(B_1)]^p / (\sup_{x \in S} \mathbb{E}_x l(B_1))^p$. □

Conclusion

In this paper, we extended the analysis of the generalization ability of MV-ERM methods to the situation where data are drawn from a (possibly nonstationary) subgeometrically ergodic Markov chains by means of the regenerative method. In particular, a novel maximal concentration inequality is established in this context. We established the generalization bound for regenerative Markov chains and showed how it can be easily extended to a Harris recurrent case.The present study paves the way for extending the (non-asymptotic) validity framework of machine-learning algorithms to situations where training data exhibit a Markovian dependence structure, or are drawn from a (pseudo-) regenerative process more generally, see [15].

Acknowledgement

This work was supported as part of the Investissement davenir, project reference ANR-11-LABX-0056-LMH and by the Polish National Science Centre NCN (grant No. UMO-2016/23/N/ST1/01355) and (partly) by the Ministry of Science and Higher Education.

References

[1] T.M. Adams and A.B. Nobel. Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling. *Annals of Probability*, 38:1345–1367, 2010.

[2] P. Alquier and O. Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18:883–913, 2012.

[3] V. de la Pena and E. Giné. *Decoupling: from Dependence to Independence*. Springer, 1999.

[4] J. Di and E. Kolaczyk. Complexity-penalized estimation of minimum volume sets for dependent data. *Journal of Multivariate Analysis*, 101(9):1910–1926, 2004.

[5] J.H.J. Einmahl and D.M. Mason. Generalized quantile process. *The Annals of Statistics*, 20:1062–1078, 1992.

[6] S. Hanneke. Learning whenever learning is possible: Universal learning under general stochastic processes. *arXiv:1706.01418*, 2017.

[7] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1996.

[8] E. Nummelin. A splitting technique for Harris recurrent chains. *Z. Wahrsch. Verw. Gebiete*, 43:309–318, 1978.

[9] V.V. Petrov. *Limit theorems of probability theory: sequences of independent random variables*. Oxford studies in probability. Clarendon Press, 1995.

[10] D. Pollard. *Convergence of Stochastic Processes*. Springer Series in Statistics, 1984.

[11] W. Polonik. Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 69(1):1–24, 1997.

[12] D. Revuz. *Markov Chains*. 2nd edition, North-Holland, 1984.

[13] C. Scott and R. Nowak. Learning minimum volume sets. *Journal of Machine Learning Research*, 7:665–704, 2006.

[14] I. Steinwart and A. Christmann. Fast learning from non-i.i.d. observations. *NIPS*, pages 1768–1776, 2009.

[15] H. Thorisson. *Coupling, Stationarity and Regeneration*. Springer, 2000.

[16] P.K. Tuominen and R. Tweedie. Subgeometric rates of convergence of f-ergodic Markov chains. *Advances in Applied Probability*, 26:775–798, 1994.