Estimating the Mixing Coefficients of Geometrically Ergodic Markov Processes

Steffen Grünewälder
gruenew@gmail.com

Azadeh Khaleghi
azadeh.khaleghi@ensae.fr

Abstract

We propose methods to estimate the individual β-mixing coefficients of a real-valued geometrically ergodic Markov process from a single sample-path $X_0, X_1, \dots, X_n$. Under standard smoothness conditions on the densities, namely, that the joint density of the pair (X_0, X_m) for each m lies in a Besov space $B^s_{1,\infty}(\mathbb{R}^2)$ for some known $s > 0$, we obtain a rate of convergence of order $\mathcal{O}(\log(n) n^{-[s]/(2[s]+2)})$ for the expected error of our estimator in this case[1]. We complement this result with a high-probability bound on the estimation error, and further obtain analogues of these bounds in the case where the state-space is finite. Naturally no density assumptions are required in this setting; the expected error rate is shown to be of order $\mathcal{O}(\log(n) n^{-1/2})$.

1 Introduction

Temporal dependence in time-series can be quantified via various notions of *mixing*, which capture how events separated over time may depend on one another. The dependence between the successive observations in a stationary sequence implies that the sequence contains less *information* as compared to an i.i.d. sequence with the same marginal distribution. This can negatively affect the statistical guarantees for dependent samples. In fact, various mixing coefficients explicitly appear in the concentration inequalities involving dependent and functions of dependent sequences, making them looser than their counterparts derived for i.i.d. samples, see, e.g. (Ibragimov, 1962; Viennet, 1997; Rio, 1999; Samson, 2000; Rio, 2000; Dedecker and Prieur, 2005; Bertail et al., 2006; Bradley, 2007; Kontorovich and Ramanan, 2008; Bosq, 2012) for a non-exhaustive list of such results. Thus, in order to be able to use these inequalities in finite-time analysis, one is often required to assume known bounds on the mixing coefficients of the processes. Moreover, the quality of the assumed upper-bounds on the mixing coefficients directly translates to the strength of the statistical guarantees involving the sequence at hand. Therefore, one way to obtain strong statistical guarantees for dependent data, is to first estimate the mixing coefficients from the samples, and then plug in the estimates (as opposed to the pessimistic upper-bounds) in the appropriate concentration inequalities. Estimating the mixing coefficients can more generally lead to a better understanding of the dependence structure in the sequence.

In this paper, we study the problem of estimating the β-mixing coefficients of a real-valued Markov chain from a finite sample-path, in the case where the process is stationary and geometrically ergodic. We start by recalling the relevant concepts.

β-mixing coefficients. Let $(\Omega, \mathfrak{F}, \mu)$ be a probability space. The β-dependence $\beta(\mathcal{U}, \mathcal{V})$ between $\mathcal{U}$ and $\mathcal{V}$ is defined as follows. Let $\iota(\omega) \mapsto (\omega, \omega)$ be the injection map from $(\Omega, \mathfrak{F})$ to $(\Omega \times \Omega, \mathcal{U} \otimes \mathcal{V})$, where $\mathcal{U} \otimes \mathcal{V}$ is the product sigma algebra generated by $\mathcal{U} \times \mathcal{V}$. Let $\mu_\otimes$ be the probability measure defined on $(\Omega \times \Omega, \mathcal{U} \otimes \mathcal{V})$

[1]We use $[s]$ to denote the integer part of the decomposition $s = [s] + \{s\}$ of $s \in (0, \infty)$ into an integer term and a *strictly positive* remainder term $\{s\} \in (0, 1]$.

obtained as the pushforward measure of μ under ι. Let $\mu_{\mathcal{U}}$ and $\mu_{\mathcal{V}}$ denote the restrictions of μ to $\mathcal{U}$ and $\mathcal{V}$ respectively. Then

$$\beta(\mathcal{U},\mathcal{V}) := \sup_{W\in\sigma(\mathcal{U}\times\mathcal{V})} |\mu_{\otimes}(W) - \mu_{\mathcal{U}}\times\mu_{\mathcal{V}}(W)| \tag{1}$$

where $\mu_{\mathcal{U}}\times\mu_{\mathcal{V}}$ is the product measure on $(\Omega\times\Omega,\mathcal{U}\otimes\mathcal{V})$ obtained from $\mu_{\mathcal{U}}$ and $\mu_{\mathcal{V}}$. This leads to the sequence $\boldsymbol{\beta} := \langle\beta(m)\rangle_{m\in\mathbb{N}}$ of β-mixing coefficients of a process $\mathbf{X}$, where $\beta(m)$ is given by $\sup_{j\in\mathbb{N}}\beta(\sigma(\{X_t : 1\le t\le j\}),\sigma(\{X_t : t\ge j+m\}))$. A stochastic process is said to be β-mixing or absolutely regular if $\lim_{m\to\infty}\beta(m)=0$.

Geometrically ergodic Markov chains. In this paper, we are concerned with real-valued stationary Markov processes that are *geometrically ergodic*. Recall that a process is stationary if for every $m,\ell\in\mathbb{N}$ the marginal distribution on $\mathbb{R}^m$ of $(X_{1+\ell},\dots,X_{m+\ell})$ is the same as that of $(X_1,\dots,X_m)$. In the case of stationary Markov processes, the β-mixing coefficient $\beta(m),\ m\in\mathbb{N}$ can be simplified to the β-dependence between the σ-algebras generated by X_1 and X_m respectively, i.e. $\beta(m)=\beta(\sigma(X_1),\sigma(X_m))$ (Bradley, 2007, vol. 1 pp. 206). A stationary Markov process is said to satisfy "geometric ergodicity" if there exists Borel functions $f:\mathbb{R}\to(0,\infty)$ and $c:\mathbb{R}\to(0,\infty)$ such that for ρ-a.e. $x\in\mathbb{R}$ and every $m\in\mathbb{N}$, it holds that $\sup_{B\in\mathfrak{B}(\mathbb{R})}|p_m(x,B)-\rho(B)|\le f(x)e^{-c(x)m}$ where $p_m(x,B)$ defined for $x\in\mathbb{R}$ and $B\in\mathfrak{B}(\mathbb{R})$ is the regular conditional distribution of X_m given X_1 and ρ denotes the marginal distribution of X_1 (Bradley, 2007, vol. 2 Definition 21.18 pp 325). It is well-known - see, e.g. (Bradley, 2007, vol. 2 Theorem 21.19 pp. 325) - that a stationary, geometrically ergodic Markov process is absolutely regular with $\beta(m)\to 0$ *at least exponentially fast* as $m\to 0$. This means that in this case the process has β-mixing coefficients of the form $\beta(m)\le\eta e^{-\gamma m}$ for some $\eta,\ \gamma\in(0,\infty)$ and all $m\in\mathbb{N}$.

Overview of the main results. Our first result involves the estimation of $\beta(m)$ for each $m=1,2,\dots,$ of a real-valued geometrically ergodic Markov chain from a finite sample path $X_1,\dots,X_n$. Our main assumption in this case is that the joint density f_m of the pair (X_0,X_m) lies in a Besov space $B^s_{1,\infty}(\mathbb{R}^2)$; roughly speaking, this implies that f_m has $[s]$ many weak derivatives. As discussed above, for a geometrically ergodic Markov chain, $\beta(m)$ is of the form $\eta^\star e^{-\gamma^\star m}$; we assume the true parameters $\eta^\star$ and $\gamma^\star$ to be unknown. Given (potentially loose) upper-bounds η and γ on $\eta^\star$ and $\gamma^\star$ we show in Theorem 3 that

$$\mathbb{E}|\beta(m)-\widehat{\beta}_N(m)|\in\mathcal{O}(\log(n)n^{-\frac{[s]}{2[s]+2}}),\ \text{for all } m\lesssim(\log n)/\gamma$$

where, $\widehat{\beta}_N$ is given by (6) with $N\approx\gamma n/\log n$ (given in Condition 2). Moreover, there exists a constant $\zeta>0$ such that with probability $1-\zeta\log(n)n^{-\frac{[s]}{2[s]+2}}$ it holds that

$$|\beta(m)-\widehat{\beta}_N(m)|\in\mathcal{O}(\log^2(n)n^{-\frac{[s]}{2[s]+2}}).$$

The constants hidden in the $\mathcal{O}$-notation are included in the full statement of the theorem. An important observation is that neither η nor γ affect the rate of convergence. However, a factor $1/\gamma$ appears in the constant, and a pessimistic upper-bound on the mixing coefficients (i.e. a small γ) can lead to a large constant in the bound on the estimation error. Theorems 4 and 5 are concerned with a different setting where the state-space of the Markov chains is finite. In this case, we do not require any assumptions on the smoothness of the densities, and the rates obtained here match that provided in Theorem 3 if we let $s\to\infty$. For the estimate $\widehat{\beta}_N(m)$ given by (8) with $N\approx(\log n)/\gamma$ and every $m\lesssim(\log n)/\gamma$ we have,

$$\mathbb{E}|\widehat{\beta}_N(m)-\beta(m)|\lesssim\frac{|\mathscr{X}|^2}{\gamma}\log(n)n^{-1/2}. \tag{2}$$

Moreover, $\Pr(|\widehat{\beta}_N(m)-\beta(m)|\ge\epsilon)\lesssim|\mathscr{X}|^2n^{-1/2}+|\mathscr{X}|^2\exp\left(-\frac{\gamma n\epsilon^2}{|\mathscr{X}|^4\log n}\right)$ for $\epsilon>0$. We refer to the statement of Theorem 4 for the explicit constants. Observe that we have a factor $|\mathscr{X}|^2/\gamma$ in (2). In other words, γ has the same effect here as in the bound provided in Theorem 3, and the constant increases quadratically in the size of the state-space. In this setting, apart from estimating the individual mixing coefficients we can also simultaneously estimate $\beta(m)$ for m up to some $k^\dagger \lesssim (\log n)/\gamma$ (see the full statement of Theorem 5 for a specification). The analysis relies on a VC-argument in place of a union bound which leads to tighter error bounds. We obtain,

$$\mathbb{E}(\sup_{m \leq k^\dagger} |\widehat{\beta}_N(m) - \beta(m)|) \lesssim \frac{1}{\gamma} \log(n|\mathscr{X}|)|\mathscr{X}|^2 n^{-1/2},$$

and, $\Pr(\sup_{m \leq k^\dagger} |\widehat{\beta}_N(m) - \beta(m)| \geq \epsilon) \lesssim |\mathscr{X}|^2 \log(n|\mathscr{X}|)n^{-1/2} + |\mathscr{X}|^2 \exp\left(-\frac{\gamma n \epsilon^2}{|\mathscr{X}|^4 \log n}\right)$ for all $\epsilon > 0$. Note that the parameter $k^\dagger$ does not explicitly appear on the right hand side of the above bounds as it has already been substituted for in the calculation.

Related literature. Research on the direct estimation of the mixing coefficients is relatively scarce. In the asymptotic regime, Nobel (2006) used hypothesis testing to give asymptotically consistent estimates of the polynomial decay rates for covariance-based mixing conditions. Khaleghi and Lugosi (2023) proposed asymptotically consistent estimators of the α-mixing and β-mixing coefficients of a stationary ergodic process from a finite sample-path. Since in general, rates of convergence are non-existent for stationary ergodic processes (see, e.g. Shields (1996)), their results necessarily remain asymptotic and no rates of convergence can be obtained. An attempt at estimating β-mixing coefficients has also been made by McDonald et al. (2015). Despite our best attempts, we have been unable to verify some of their main claims and have particular reservations about the validity of their rates. More specifically, their main theorem (Theorem 4) suggests a rate of convergence of order $\log(n)n^{-1/2}$ for their estimator, independently of the dimension of the state-space and under the most minimal smoothness assumptions on the densities. Given that under these conditions a density estimator is known to have a dimension-dependent rate of about $n^{-1/(2+d)}$ even when the samples are iid (Giné and Nickl, 2021, pp. 404), it is highly unlikely that a dimension-independent rate would be achievable for an estimator of the β-mixing coefficient. We would like to point out that an interesting body of work exists for a different, yet related problem, concerning the estimation of the mixing times of finite-state Markov chains (Hsu. et al., 2019; Wolfer and Kontorovich, 2019; Wolfer, 2020). We believe that the techniques developed in this line of work may have strong links to the estimation of the α-mixing (as opposed to the β-mixing) coefficients of finite-state Markov chains.

2 Preliminaries

In this section we introduce notation and provide some basic definitions. We denote the non-negative integers by $\mathbb{N} := \{0, 1, 2, \dots\}$. If $s \in (0, \infty)$, then we let $s = [s] + \{s\}$ be decomposed into its integer part $[s] \in \mathbb{N}$ and a *strictly positive* remainder term $\{s\} \in (0, 1]$. In particular, if $s = i$ for any $i > 0 \in \mathbb{N}$ then, $[s] = i - 1$ and $\{s\} = 1$. As part of our analysis in Section 3.1, we impose classical density assumptions on certain finite-dimensional marginals of the Markov chains considered. The densities satisfy standard smoothness conditions as controlled by the parameters of appropriate Besov spaces.

Besov Spaces $B^s_{p,\infty}(\mathbb{R}^d)$. For an arbitrary function $f : \mathbb{R}^d \to \mathbb{R}$ and any vector $h \in \mathbb{R}^d$ let $\Delta_h f(x) := f(x + h) - f(x)$ be the first-difference operator, and obtain higher-order differences inductively by $\Delta^r_h f := (\Delta_h \circ \Delta^{r-1}_h) f$ for $r = 2, 3, \dots$. Denote by $L_p(\mathbb{R}^d)$, $p \geq 1$ the L_p space of functions $f : \mathbb{R}^d \to \mathbb{R}$. For $s > 0$ the Besov space $B^s_{p,\infty}(\mathbb{R}^d)$ is defined as

$$B^s_{p,\infty}(\mathbb{R}^d) := \{f \in L_p(\mathbb{R}^d) : \|f\|_{B^s_{p,\infty}(\mathbb{R}^d)} < \infty\} \tag{3}$$

with Besov norm $\|f\|_{B^s_{p,\infty}(\mathbb{R}^d)} := \|f\|_1 + \sup_{0<t<\infty} t^{-s} \sup_{|h| \leq t} \|\Delta^r_h f\|_1$ where $|v| := \sum_{i=1}^d |v_i|$ for $v = (v_1, \dots, v_d) \in \mathbb{R}^d$, and r is any integer such that $r > s$ (Bennett and Sharpley, 1988). Denote by

$W_p^r(\mathbb{R}^d),\ r \in \mathbb{N}$ the Sobolev space of functions $f : \mathbb{R}^d \to \mathbb{R}$. We rely on the following interpolation result concerning $W_p^r(\mathbb{R}^d)$ and $B^s_{p,\infty}(\mathbb{R}^d)$.

Remark 1. *As follows from (Bennett and Sharpley, 1988, Proposition 5.1.8 and Theorem 5.4.14) for any* $r_0,\ r_1 \in \mathbb{N}$ *and* $\varsigma := (1-\theta)r_0 + \theta r_1,\ \theta \in (0,1)$ *it holds that*

$$W_p^{r_0}(\mathbb{R}^d) \cap W_p^{r_1}(\mathbb{R}^d) \hookrightarrow B^{\varsigma}_{p,\infty}(\mathbb{R}^d) \hookrightarrow W_p^{r_0}(\mathbb{R}^d) + W_p^{r_1}(\mathbb{R}^d). \tag{4}$$

If $r_0 < r_1$*, then the left and right hand sides of* (4) *reduce to* $W_p^{r_1}(\mathbb{R}^d)$ *and* $W_p^{r_0}(\mathbb{R}^d)$ *respectively. Furthermore, in this case we have* $\|f\|_{B^{\varsigma}_{p,\infty}(\mathbb{R}^d)} \geq \|f\|_{W_p^{r_0}(\mathbb{R}^d)}$*. To see this consider the* K*-functional*

$$K(f, t; W_p^{r_0}(\mathbb{R}^d), W_p^{r_1}(\mathbb{R}^d)) := \inf\{\|f_0\|_{W_p^{r_0}(\mathbb{R}^d)} + t\|f_1\|_{W_p^{r_1}(\mathbb{R}^d)} : f = f_0 + f_1\}$$

and observe that by (Bennett and Sharpley, 1988, Theorem 5.4.14 and Definition 5.1.7) we have $\|f\|_{B^{\varsigma}_{p,\infty}(\mathbb{R}^d)} = \sup_{t>0} t^{-\theta} K(f, t; W_p^{r_0}(\mathbb{R}^d), W_p^{r_1}(\mathbb{R}^d)) \geq K(f, 1; W_p^{r_0}(\mathbb{R}^d), W_p^{r_1}(\mathbb{R}^d)) \geq \inf\{\|f_0\|_{W_p^{r_0}(\mathbb{R}^d)} + \|f_1\|_{W_p^{r_0}(\mathbb{R}^d)} : f = f_0 + f_1\} \geq \|f\|_{W_p^{r_0}(\mathbb{R}^d)}$*. In particular, consider the Besov space* $B^s_{1,\infty}(\mathbb{R}^d)$ *for some* $s \in (0,\infty)$ *and take* $\theta = \{s\}/2$*,* $r_0 = [s]$ *and* $r_1 = [s] + 2$ *for the above convex combination. Then* $\|f\|_{B^s_{1,\infty}(\mathbb{R}^d)} \geq \|f\|_{W_1^{[s]}(\mathbb{R}^d)}$*.*

3 Problem formulation and main results

We are given a sample $X_0, X_1, \ldots, X_{n-1}$ generated by a stationary geometrically ergodic Markov chain taking values in some $\mathscr{X} \subseteq \mathbb{R}$. As discussed in Section 2 such a process is known to have a sequence of β-mixing coefficients of the form

$$\beta(m) \leq \eta^\star e^{-\gamma^\star m},\ m \in \mathbb{N}$$

for some unknown constants $\eta^\star,\ \gamma^\star \in (0,\infty)$. The mixing coefficient $\beta(m)$ and its rate are unknown, and our objective is to estimate its rate parameters $\eta^\star$ and $\gamma^\star$. We focus on two different settings, depending on the state-space. First, in Section 3.1, we consider the case where the process is real-valued and its one and two-dimensional marginals have densities with respect to the Lebesgue measure. Next, in Section 3.2, we consider the case where $\mathscr{X}$ is finite. In this setting we do not require any density assumptions, and are able to control the estimation error simultaneously for multiple values of m. This is stated in Theorem 5. The following notation is used in both settings. For each $m \in \mathbb{N}$, denote by P_m the joint distribution of the pair (X_0, X_m) so that for each $U \in \mathfrak{B}(\mathscr{X}^2)$ we have $\Pr(\{(X_0, X_m) \in U\}) = P_m(U)$. By stationarity, $\Pr(\{(X_t, X_{m+t}) \in U\}) = P_m(U),\ t \in \mathbb{N}$.

3.1 Real-valued state-space

We start by considering the case where $\mathscr{X}$ is any subset of $\mathbb{R}$. We assume that P_m has a density $f_m : \mathbb{R}^2 \to \mathbb{R}^2$ with respect to the Lebesgue measure λ_2 on $\mathbb{R}^2$. Denote by P_0 the marginal distribution of $X_t,\ t \in \mathbb{N}$ whose density $f_0 : \mathbb{R} \to \mathbb{R}$ with respect to the Lebesgue measure λ on $\mathbb{R}$ can be obtained as $f_0(x) = \int_{\mathbb{R}} f_m((x,y))d\lambda(y)$ for $x \in \mathbb{R}$. It follows that $\beta(m) = \frac{1}{2}\int_{\mathbb{R}^2} |f_m - f_0 \otimes f_0| d\lambda_2$. For some fixed $k \in m+1, \ldots, \lfloor n/8 \rfloor$ let $N = N(k,n) := \lfloor \frac{n-k}{2(k+1)} \rfloor$ and define the sequence of tuples $Z_i = (X_{2i(k+1)}, X_{2i(k+1)+m}),\ i = 0, 1, 2, \ldots, N-1$. Define the Kernel Density Estimator (KDE) of f_m as

$$\widehat{f}_{m,N}(z) = \frac{1}{N h_N^2} \sum_{i=1}^{N} K\left(\frac{z - Z_i}{h_N}\right) \tag{5}$$

with kernel $K : \mathbb{R}^2 \to \mathbb{R}$ and bandwidth $h_N > 0$. Marginalizing we obtain an empirical estimate of f_0, i.e. $\widehat{f}_{0,N}(x) := \int_{\mathbb{R}} \widehat{f}_{m,N}(x,y)d\lambda(y)$. We define an estimator of $\beta(m)$ as

$$\widehat{\beta}_N(m) = \frac{1}{2}\int_{\mathbb{R}^2} |\widehat{f}_{m,N} - \widehat{f}_{0,N} \otimes \widehat{f}_{0,N}| d\lambda_2 \tag{6}$$

where $\otimes$ denotes the tensor product. Note that to simplify notation in (5) and (6), we have omitted the dependence of N on the choice k. An optimal value for k, denoted, k^* is provided in Condition 2. In our analysis we make standard assumptions (see, Condition 2 below) about the smoothness of f_m and the order of the kernel K in (5). Recall that a bivariate kernel is said to be of order ℓ if

1. $c_\ell(K) := \displaystyle\sum_{\substack{i,j\in\mathbb{N}\\ i+j=\ell}} \int_{z=(z_1,z_2)\in\mathbb{R}^2} |z_1|^i |z_2|^j |K(z)| d\lambda_2(z) < \infty$
2. $\displaystyle\int_{z\in\mathbb{R}^2} K(z) d\lambda_2(z) = 1$ and $\displaystyle\int_{z=(z_1,z_2)\in\mathbb{R}^2} z_1^i z_2^j K(z) d\lambda_2(z) = 0$ for all $i,\ j\in\mathbb{N},\ i+j<\ell$

Condition 2. *The density $f_m \in B^s_{1,\infty}(\mathbb{R}^2)$ for some $s>1$ and $\|f_m\|_{B^s_{1,\infty}(\mathbb{R}^2)} \leq \Lambda$ for some $\Lambda\in(0,\infty)$. It is further assumed that the ℓ^{th} moments of the pair (X_1, X_m) are finite for $\ell = 1,\dots,\lceil s\rceil$, and $\int_{\mathbb{R}^2} f(z)(1+\|z\|^2)d\lambda_2(z)$, $\int_{\mathbb{R}^2} K^2(z)(1+\|z\|^2)d\lambda_2(z) < \infty$. While the parameters $\eta^\star$ and $\gamma^\star$ are unknown, some lower-bound $\gamma \leq \gamma^\star$ and some upper-bound $\eta \geq \eta^\star$ are given. The estimator $\widehat{\beta}_N(m)$ given by (6) for $m = 1,\dots,k^*$ is obtained via*

i. a convolution kernel K of order $[s]$ such that $c_0 := \int_{\mathbb{R}^2} |K(z)| d\lambda_2(z) < \infty$.

ii. and a bandwidth of the form $h_N = (c\Lambda)^{-\frac{[s]}{[s]+1}} \left(\frac{[s]-1}{2}\right)^{-\frac{[s]}{2[s]+2}} n^{-\frac{1}{2[s]+2}}$

where, $c := \dfrac{c_{[s]}(K)}{[s]!}$ and $N = \lfloor \frac{n-k^*}{2(k^*+1)} \rfloor$ with $k^* = \frac{1}{\gamma}\left(\log\frac{\gamma\eta}{8C} + (\frac{3[s]+2}{2[s]+2})\log n\right)$, and $C := (2+c_0)(L_1)^{\frac{[s]}{[s]+1}}(c\Lambda)^{\frac{1}{[s]+1}}$ with $L_1^2 := \frac{2}{s-1}\int_{\mathbb{R}^2} f(z)(1+\|z\|^2)d\lambda_2(z)\int_{\mathbb{R}^2} K^2(z)(1+\|z\|^2)d\lambda_2(z)$.

We are now in a position to state our main result, namely, Theorem 3 below which provides bounds on the estimation error of $\widehat{\beta}_N$ given by (6), when the assumptions stated in Condition 2 are satisfied. Note that this condition is fulfilled by a number of standard models. For instance, consider the stationary, geometrically ergodic $AR(1)$ model $X_{t+1} = aX_t + \epsilon_t,\ t\in\mathbb{N}$, where $\epsilon_t \sim \mathcal{N}(0,\sigma^2)$ for some $\sigma>0$ and where $X_0 \sim \mathcal{N}(0, \sigma^2/(1-a^2))$ for some $a\in\mathbb{R}$ with $|a|<1$. All of the finite-dimensional marginals of this process are Gaussian; in particular, its marginal and joint densities f_0 and $f_m,\ m=1,2,\dots$, being infinitely differentiable, lie in any of the Besov spaces that we consider in this paper.

Theorem 3. *Under the assumptions stated and with the parameters defined in Condition 2, for each $m \in 1,\dots,k^*$ we have*

$$\mathbb{E}|\beta(m) - \widehat{\beta}_N(m)| \leq \frac{8Cn^{-\frac{[s]}{2[s]+2}}}{\gamma}\left(1 + \log\frac{\gamma\eta}{8C} + 2\log n\right)$$

Moreover, with probability $1 - \left(2 + 8C\log\left(\frac{\gamma\eta}{8C} + \frac{3}{2}\log n\right)\right)n^{-\frac{[s]}{2[s]+2}}$ it holds that

$$|\beta(m) - \widehat{\beta}_N(m)| \leq 64\big(1 + \frac{c_0}{2}\big)(C_1 + C_2\log(n) + \frac{6\|K\|_1}{\gamma}\log^2(n))n^{-\frac{[s]}{2[s]+2}}.$$

where

$$C_1 = \frac{1}{\gamma}\log\left(\frac{\gamma\eta}{8C}\right)\left(3(L_1)^{\frac{[s]}{[s]+1}}(c\Lambda)^{\frac{1}{[s]+1}} + \Lambda(c\Lambda)^{-\frac{[s]^2}{[s]+1}}\left(\frac{[s]-1}{2}\right)^{-\frac{[s]^2}{2[s]+2}}\right),$$

$$C_2 = \frac{4\|K\|_1}{\gamma}\log\left(\frac{\gamma\eta}{8C}\right) + \frac{3}{2\gamma}\left(3(L_1)^{\frac{[s]}{[s]+1}}(c\Lambda)^{\frac{1}{[s]+1}} + \Lambda(c\Lambda)^{-\frac{[s]^2}{[s]+1}}\left(\frac{[s]-1}{2}\right)^{-\frac{[s]^2}{2[s]+2}}\right).$$

See Section 4 for a proof.

3.2 Finite state-space

In the special case where the state-space $\mathscr{X}$ of the Markov chain is finite, we can relax the density assumptions and obtain an empirical estimate of β by counting frequencies. More specifically, in this case, for each $t \in \mathbb{N}$ the σ-algebra $\sigma(X_t)$ is completely atomic with atoms $\{X_t = s\}, s \in \mathscr{X}$. Therefore, by (Bradley, 2007, vol. Proposition 3.21 pp. 88) we have

$$\beta(m) = \sum_{u \in \mathscr{X}} \sum_{v \in \mathscr{X}} |P_m(\{(u,v)\}) - P_0(\{u\})P_0(\{v\})| \tag{7}$$

where as before, P_m and P_0 are the joint and the marginal distributions of (X_0, X_m) and X_0 respectively. Given a sample $X_0, \ldots, X_{n-1}$, we can obtain an empirical estimate of $\beta(m)$ in (7) as follows. Fix a lag of length $k \in 1, \ldots, n$ (an optimal value for which will be specified in Proposition 4). Define the sequence of tuples $Z_i = (X_{2ki}, X_{2k(i+1)})$, $i = 0, 2, \ldots, N-1$, with $N = N(k,n) := \lfloor \frac{n-k}{2(k+1)} \rfloor$. For each pair $(u,v) \in \mathscr{X}^2$ let $\widehat{P}_{m,N}((u,v)) = \frac{1}{N}\sum_{i=0}^{N-1} \mathbf{1}_{\{(u,v)\}}(Z_i)$. Similarly, for each $u \in \mathscr{X}$ we can obtain an empirical estimate of $P_0(u)$ as $\widehat{P}_{0,N}(u) := \frac{1}{2N}\sum_{i=0}^{2N} \mathbf{1}_{\{u\}}(X_{ki})$. Define

$$\widehat{\beta}_N(m) := \sum_{u \in \mathscr{X}} \sum_{v \in \mathscr{X}} |\widehat{P}_{m,N}(\{(u,v)\}) - \widehat{P}_{0,N}(\{u\})\widehat{P}_{0,N}(\{v\})| \tag{8}$$

Theorem 4. *Consider a sample of length $n \in \mathbb{N}$ of a stationary, geometrically ergodic Markov chain with finite state-space $\mathscr{X}$. Define $N(k,n) = \lfloor \frac{n-k}{2(k+1)} \rfloor$, $n, k \in \mathbb{N}$. Let $k^\star := \frac{1}{\gamma}\log\left(\frac{\eta\gamma n^{3/2}}{\sqrt{8}|\mathscr{X}|^2}\right)$ and $N = N(k^\star, n)$, $n \geq \max\left\{2|\mathscr{X}|^3\left(\frac{e^\gamma}{\eta\gamma}\right)^{2/3}, \left(\frac{\eta\gamma}{|\mathscr{X}|^2}\right)^{2/3}\right\}$. For every $m = 1, \ldots, k^\star$ and $\widehat{\beta}_N(m)$ given by* (8) *we have,*

$$\mathbb{E}|\widehat{\beta}_N(m) - \beta(m)| \leq \frac{\sqrt{32}|\mathscr{X}|^2 n^{-1/2}}{\gamma}\left(1 + \log\left(\frac{\eta\gamma}{\sqrt{8}|\mathscr{X}|^2}\right) + \frac{3}{2}\log n\right)$$

Moreover, for $\epsilon > 0$ it holds that

$$\Pr(|\widehat{\beta}_N(m) - \beta(m)| \geq \epsilon) \leq \frac{\sqrt{2}|\mathscr{X}|^2 n^{-1/2}}{\log(\frac{\eta\gamma n^{3/2}}{\sqrt{8}|\mathscr{X}|^2})} + 4|\mathscr{X}|^2 \exp\{-\frac{\gamma n \epsilon^2}{48|\mathscr{X}|^4 \log n}\}$$

The proof is provided in Section 4.

In this setting, we are also able to simultaneously control the estimation error for all $m = 1, \ldots, k^\dagger$ where $k^\dagger$ is specified in the statement of Theorem 5. The proof relies on a VC argument which helps replace a factor of $k^\dagger$ (which would have otherwise been deduced from a union bound) with a factor of $\log k^\dagger$.

Theorem 5. *Consider a sample of length $n \in \mathbb{N}$ of a stationary, geometrically ergodic Markov chain with finite state-space $\mathscr{X}$. Define $N(k,n) = \lfloor \frac{n-k}{2(k+1)} \rfloor$, $n, k \in \mathbb{N}$. Let $N = N(k^\dagger, n)$ with $k^\dagger := \frac{1}{\gamma}\log\left(\frac{\eta\gamma n^{3/2}}{8\sqrt{2}|\mathscr{X}|^2\log(n|\mathscr{X}|)}\right)$ and $n \geq \max\left\{\frac{8\sqrt{2}|\mathscr{X}|^3 e^\gamma}{\eta\gamma}, \frac{\eta\gamma}{8\sqrt{2}|\mathscr{X}|}\right\}$. For $\widehat{\beta}_N$ given by* (8) *we have,*

$$\mathbb{E}[\sup_{m \in 1,\ldots,k^\dagger} |\widehat{\beta}_N(m) - \beta(m)|] \leq \frac{8\sqrt{2}|\mathscr{X}|^2 n^{-1/2}\log(n|\mathscr{X}|)}{\gamma}\left(1 + \log\left(\frac{\eta\gamma n^{1/2}}{8\sqrt{2}|\mathscr{X}|^2\log(n|\mathscr{X}|)}\right)\right)$$

Furthermore, for $\epsilon > 0$ the probability $\Pr(\sup_{m \in 1,\ldots,k^\dagger} |\widehat{\beta}_N(m) - \beta(m)| \geq \epsilon)$ is at most

$$\frac{4\sqrt{2}|\mathscr{X}|^2\log(n|\mathscr{X}|)n^{-1/2}}{\log(\frac{\eta\gamma n^{3/2}}{8\sqrt{2}|\mathscr{X}|^2})} + 16|\mathscr{X}|^2\log\left(\frac{3|\mathscr{X}|}{2\gamma}\log(\eta^{2/3}\gamma^{2/3}n)\right)\exp\left\{-\frac{\gamma n\epsilon^2}{3072|\mathscr{X}|^4\log n}\right\}$$

The proof is provided in Section 4.

4 Proofs

In this section we provide a proof for our theorems. A common ingredient in all three proofs is a coupling argument for time-series, which allows one to move from dependent samples to independent blocks. This is facilitated by Lemma 6 below, which is a standard result based on, commonly used in the analysis of dependent time-series, see e.g. (Levental, 1988; Yu, 1994; Arcones and Yu, 1994). For completeness, we provide a proof of this lemma, which in turns relies on a coupling Lemma of Berbee (1979) stated below.

Lemma 6. *Let $X_i,\ i \in \mathbb{N}$ be a stationary sequence of random variables with β-mixing coefficients $\beta(j),\ j \in \mathbb{N}$. For a fixed $k,\ \ell \in \mathbb{N}$ let $Y_i = X_{i(k+\ell)}, \ldots, X_{ik+(i+1)\ell}$ for $i \in \mathbb{N}$. There exists a sequence of independent random variables $Y_i^*,\ i \in \mathbb{N}$ taking values in $\mathbb{R}^\ell$ and have the same distribution as Y_i such that for every $i \in \mathbb{N}$ we have,*

$$\Pr(Y_i^* \neq Y_i) \leq \beta(k).$$

Lemma 7 (Berbee (1979))**.** *Let X and Y be two random variables taking values in Borel spaces $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively. Denote by U a random variable uniformly distributed over $[0,1]$, which is independent of (X,Y). There exists a random variable $Y^* = g(X,Y,U)$ where $g : \mathcal{S}_1 \times \mathcal{S}_2 \times [0,1] \to \mathcal{S}_2$ such that Y^* is independent of X and has the same distribution as Y, and that $\Pr(Y^* \neq Y) = \beta(\sigma(X), \sigma(Y))$.*

Proof of Lemma 6. Let $U_j,\ j \in \mathbb{N}$ be a sequence of i.i.d. random variables uniformly distributed over $[0,1]$ such that each U_j is independent of $\sigma(\{Y_i : i \in \mathbb{N}\})$. Set $Y_0^* = Y_0$. By Lemma 7 there exists a random variable $Y_1^* = g_1(Y_0^*, Y_1, U_1)$ where g_1 is a measurable function from $\mathbb{R}^\ell \times \mathbb{R}^\ell \times [0,1]$ to $\mathbb{R}^\ell$ such that Y_1^* is independent of Y_0^*, has the same distribution as Y_1 and $\Pr(Y_1^* \neq Y_1) = \beta(\sigma(Y_0^*), \sigma(Y_1))$. Similarly, there exists a random variable $Y_2^* = g_2((Y_0^*, Y_1^*), Y_2, U_2)$ where g_2 is a measurable function from $(\mathbb{R}^\ell)^2 \times \mathbb{R}^\ell \times [0,1]$ to $\mathbb{R}^\ell$ such that Y_2^* is independent of (Y_0^*, Y_1^*), has the same distribution as Y_2 and $\Pr(Y_2^* \neq Y_2) = \beta(\sigma(Y_0^*, Y_1^*), \sigma(Y_2))$. Continuing inductively in this way, at each step $j = 3, 4, \ldots$, by Lemma 7, there exists a random variable $Y_j^* = g_j((Y_0^*, Y_1^*, \ldots, Y_{j-1}^*), Y_j, U_j)$ where g_j is a measurable function from $(\mathbb{R}^\ell)^j \times \mathbb{R}^\ell \times [0,1]$ to $\mathbb{R}^\ell$ such that Y_j^* is independent of $(Y_0^*, Y_1^*, \ldots, Y_{j-1}^*)$, has the same distribution as Y_j and that $\Pr(Y_j^* \neq Y_j) = \beta(\sigma(Y_0^*, Y_1^*, \ldots, Y_{j-1}^*), \sigma(Y_j))$. It remains to show that $\beta(\sigma(Y_0^*, Y_1^*, \ldots, Y_{j-1}^*), \sigma(Y_j)) \leq \beta(k)$ for all $j \in \mathbb{N}$. To see this, first note that $Y_0^* = Y_0$ by definition, and that for each $i \in \mathbb{N}$, it holds that $Y_i^* \in \sigma((Y_0^*, Y_1^*, \ldots, Y_{i-1}^*), Y_i, U_i)$, we have

$$\sigma(Y_0^*, Y_1^*, \ldots, Y_{j-1}^*) \subseteq \mathcal{U}_j \vee \mathcal{V}_j \tag{9}$$

where $\mathcal{U}_j = \sigma(U_1, \ldots, U_{j-1})$ and $\mathcal{V}_j = \sigma(Y_0, Y_1, \ldots, Y_{j-1})$. Take any $U \in \mathcal{U}_j$ and $W \in \sigma(Y_j)$. We almost surely have,

$$\begin{aligned}
P(U \cap W|\mathcal{V}_j) &= \mathbb{E}\left(\mathbf{1}_U \mathbf{1}_W|\mathcal{V}_j\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\mathbf{1}_U \mathbf{1}_W|\mathcal{V}_j \vee \sigma(Y_j)\right)|\mathcal{V}_j\right) && \text{since } \mathcal{V}_j \subseteq \sigma(Y_j) \vee \mathcal{V}_j \\
&= \mathbb{E}\left(\mathbf{1}_W \mathbb{E}\left(\mathbf{1}_U|\mathcal{V}_j \vee \sigma(Y_j)\right)|\mathcal{V}_j\right) && \text{since } W \in \sigma(Y_j) \\
&= \mathbb{E}\left(\mathbf{1}_W \mathbb{E}\left(\mathbf{1}_U\right)|\mathcal{V}_j\right) && \text{since } \mathcal{U}_j \text{ is independent of } \sigma(Y_0, \ldots, Y_j) \\
&= P(U)P(W|\mathcal{V}_j) \\
&= P(U|\mathcal{V}_j)P(W|\mathcal{V}_j) && \text{since } \mathcal{U}_j \text{ is independent of } \mathcal{V}_j
\end{aligned}$$

Therefore, and $\mathcal{U}_j$, $\mathcal{V}_j$ and $\sigma(Y_j)$ form a Markov triplet in the sense of (Bradley, 2007, Vol. 1 Definition 7.1 pp. 205). Thus, as follows from (Bradley, 2007, Vol. 1 Theorem 7.2 pp. 205) we obtain,

$$\beta(\mathcal{U}_j \vee \mathcal{V}_j, \sigma(Y_j)) = \beta(\mathcal{V}_j, \sigma(Y_j)). \tag{10}$$

In light of (9) and (10), and noting that by construction $\beta(\mathcal{V}_j, \sigma(Y_j)) \leq \beta(k)$ we obtain

$$\beta(\sigma(Y_0^*, Y_1^*, \ldots, Y_{j-1}^*), \sigma(Y_j)) \leq \beta(\mathcal{U}_j \vee \mathcal{V}_j, \sigma(Y_j)) \leq \beta(\mathcal{V}_j, \sigma(Y_j)) \leq \beta(k).$$

$\square$

Proof of Theorem 3. Given the sample $X_0, \ldots, X_n$, consider the sequence $Z_i = (X_{2i(k+1)}, X_{2i(k+1)+m})$ with $i = 0, 1, 2, \ldots, N-1$ where $N = N(k,n) := \lfloor \frac{n-k}{2(k+1)} \rfloor$ for some fixed $k \in m+1, \ldots, \lfloor n/8 \rfloor$. (As part of the proof, we propose an optimal choice for k, see (24).) Enlarge Ω if necessary in order for Lemma 6 to be applicable. As follows from Lemma 6 there exists a sequence of independent random variables $Z_i^*,\ i = 0, 1, \ldots, N-1$ each of which takes value in $\mathbb{R}^2$ and has the same distribution as $Z_i,\ i = 0, 1, \ldots, N-1$, with the additional property that

$$\Pr\left(\{\exists i \in 0, \ldots, N-1 : Z_i^* \neq Z_i\}\right) \leq N\beta(k) \tag{11}$$

Define the KDE of f_m through $Z_i^*,\ i = 0, \ldots, N-1$

$$\widehat{f}^*_{m,N}(z) = \frac{1}{Nh_N^2} \sum_{i=1}^{N} K\left(\frac{z - Z_i^*}{h_N}\right)$$

with the same kernel $K : \mathbb{R}^2 \to \mathbb{R}$ and bandwidth $h_N > 0$ as in (5) and let

$$\widehat{\beta}^*_N(m) = \frac{1}{2} \int_{\mathbb{R}^2} |\widehat{f}^*_{m,N} - \widehat{f}^*_{0,N} \otimes \widehat{f}^*_{0,N}| d\lambda_2 \tag{12}$$

where $\widehat{f}^*_{0,N}(x) := \int_{\mathbb{R}} \widehat{f}^*_{m,N}(x,y) d\lambda(y)$.

$$\begin{aligned}
&\|f_0 \otimes f_0 - \widehat{f}^*_{0,N} \otimes \widehat{f}^*_{0,N}\|_1 \\
&= \int_x \int_y |f_0(x)f_0(y) - \widehat{f}^*_{0,N}(x)\widehat{f}^*_{0,N}(y)| d\lambda(x) d\lambda(y) \\
&= \int_x \int_y |f_0(x)f_0(y) - \widehat{f}^*_{0,N}(x)f_0(y) + \widehat{f}^*_{0,N}(x)f_0(y) - \widehat{f}^*_{0,N}(x)\widehat{f}^*_{0,N}(y)| d\lambda(x) d\lambda(y) \\
&\leq \int_y f_0(y) \int_x |f_0(x) - \widehat{f}^*_{0,N}(x)| d\lambda(x) d\lambda(y) + \int_x |\widehat{f}^*_{0,N}(x)| \int_y |f_0(y) - \widehat{f}^*_{0,N}(y)| d\lambda(y) d\lambda(x) \\
&\leq (1 + c_0)\|f_0 - \widehat{f}^*_{0,N}\|_1
\end{aligned} \tag{13}$$

where $c_0 = \int_{\mathbb{R}^2} |K(z)| d\lambda_2(z)$ as specified in the theorem statement. It follows that

$$\begin{aligned}
|\beta(m) - \widehat{\beta}^*_N(m)| &= \frac{1}{2} \left| \int |f_m - f_0 \otimes f_0| d\lambda_2 - \int |\widehat{f}^*_{m,N} - \widehat{f}^*_{0,N} \otimes \widehat{f}^*_{0,N}| d\lambda_2 \right| \\
&\leq \frac{1}{2} \left| \int \left(|f_m - \widehat{f}^*_{m,N}| + |f_0 \otimes f_0 - \widehat{f}^*_{0,N} \otimes \widehat{f}^*_{0,N}| \right) d\lambda_2 \right| \\
&= \frac{1}{2} \int |f_m - \widehat{f}^*_{m,N}| d\lambda_2 + \frac{1}{2} \int |f_0 \otimes f_0 - \widehat{f}^*_{0,N} \otimes \widehat{f}^*_{0,N}| d\lambda_2 \\
&= \frac{1}{2} \|f_m - \widehat{f}^*_{m,N}\|_1 + \frac{1}{2} \|f_0 \otimes f_0 - \widehat{f}^*_{0,N} \otimes \widehat{f}^*_{0,N}\|_1 \\
&\leq \frac{1}{2} \left(\|f_m - \widehat{f}^*_{m,N}\|_1 + (1 + c_0)\|f_0 - \widehat{f}^*_{0,N}\|_1 \right)
\end{aligned} \tag{14}$$

where (14) follows from (13). Next, it is straightforward to check that if $f_m \in B^s_{1\infty}(\mathbb{R}^2)$ with $\|f\|_{B^s_{1\infty}(\mathbb{R}^2)} \leq \Lambda$, then $f_0 = \int_{\mathbb{R}} f_m d\lambda \in B^s_{1\infty}(\mathbb{R})$ with $\|f\|_{B^s_{1\infty}(\mathbb{R})} \leq \Lambda$. Moreover, observe that, as follows from Remark 1, for all $f \in B^s_{1\infty}(\mathbb{R})$ and $g \in B^s_{1\infty}(\mathbb{R}^2)$ we have $\|f\|_{W_1^{[s]}(\mathbb{R})} \leq \|f\|_{B^s_{1,\infty}(\mathbb{R})}$ and $\|g\|_{W_1^{[s]}(\mathbb{R}^2)} \leq \|g\|_{B^s_{1,\infty}(\mathbb{R}^2)}$. Therefore, with the choice of bandwidth h specified in the theorem statement, by (Giné and Nickl, 2021, Proposition 4.1.5 and Proposition 4.3.33) and an argument analogous to that of (Giné and Nickl, 2021, Proposition 5.1.7), we obtain,

$$\sup_{f_0 : \|f\|_{B^s_{1\infty}(\mathbb{R})} \leq \Lambda} \mathbb{E}\|\widehat{f}^*_{0,N} - f_0\|_1 \leq \widetilde{C} N^{-\frac{[s]}{2[s]+2}} \tag{15}$$

where $\widetilde{C} = 2(L_1)^{\frac{[s]}{[s]+1}}(c\Lambda)^{\frac{1}{[s]+1}}$. Similarly, by (Giné and Nickl, 2021, pp. 404) we have,

$$\sup_{f_m:\|f\|_{B^s_{1\infty}(\mathbb{R}^2)}\le\Lambda} \mathbb{E}\|\widehat{f}^*_{m,N} - f_m\|_1 \le \widetilde{C}N^{-\frac{[s]}{2[s]+2}} \tag{16}$$

Set $C := (2 + c_0)\widetilde{C}/2$. Define the event $E := \{Z^*_i = Z_i,\ i \in 0, \dots, N-1\}$. We obtain,

$$\begin{aligned}
\mathbb{E}|\beta(m) - \widehat{\beta}_N(m)| &\le \mathbb{E}|\beta(m) - \widehat{\beta}^*_N(m)| + \mathbb{E}[|\widehat{\beta}^*_N(m) - \widehat{\beta}_N(m)|\,|E^c]\Pr(E^c) &(17)\\
&\le \mathbb{E}|\beta(m) - \widehat{\beta}^*_N(m)| + 2N\beta(k) &(18)\\
&\le CN^{-\frac{[s]}{2[s]+2}} + 2N\beta(k) &(19)\\
&\le CN^{-\frac{[s]}{2[s]+2}} + 2N\eta e^{-\gamma k} &(20)\\
&\le C\left(\frac{4k}{n-4k}\right)^{\frac{[s]}{2[s]+2}} + \frac{n}{k}\eta e^{-\gamma k} &(21)\\
&\le C\left(\frac{8k}{n}\right)^{\frac{[s]}{2[s]+2}} + \frac{n}{k}\eta e^{-\gamma k} &(22)\\
&\le \frac{8Ck}{n^{\frac{[s]}{2[s]+2}}} + \eta n e^{-\gamma k} &(23)
\end{aligned}$$

where (17) follows from triangle inequality and observing that under E the estimators $\widehat{\beta}_N$ and $\widehat{\beta}^*_N$ are equal, (18) follows from (11), (19) follows from (14),(15) and (16), (20) follows from observing that $\beta(k) \le 1$ and the geometric ergodicity of the chain, and (21) and (22) follow from the definition of N and the fact that $2 \le k \le \lfloor n/8 \rfloor$. Optimizing (23) for k we obtain

$$k^\star = \frac{1}{\gamma}\left(\log\frac{\gamma\eta}{8C} + \left(\frac{3[s]+2}{2[s]+2}\right)\log n\right) \tag{24}$$

which in turn leads to

$$\mathbb{E}|\beta(m) - \widehat{\beta}^*_N(m)| \le \frac{8Cn^{-\frac{[s]}{2[s]+2}}}{\gamma}\left(1 + \log\frac{\gamma\eta}{8C} + 2\log n\right) \tag{25}$$

This completes the proof of the bound on the expected error. For the high probability bound observe that (Giné and Nickl, 2021, Theorem 5.1.13) states that for all $t > 0$,

$$\Pr\big(N\|\widehat{f}^*_{m,N} - \mathbb{E}(\widehat{f}^*_{m,N})\|_1 \ge (3/2)N\mathbb{E}\|\widehat{f}^*_{m,N} - \mathbb{E}(\widehat{f}^*_{m,N})\|_1 + \sqrt{2Nt}\|K\|_1 + t5\|K\|_1\big) \le e^{-t}. \tag{26}$$

Furthermore, $\mathbb{E}(\widehat{f}^*_{m,N}) = K_{h_N} * f_m$, where $K_h(x) = (1/h)K(x/h)$ for $h > 0, x \in \mathbb{R}^2$. Hence,

$$\|\widehat{f}^*_{m,N} - f_m\|_1 \le \|\widehat{f}^*_{m,N} - \mathbb{E}(\widehat{f}^*_{m,N})\|_1 + \|K_{h_N} * f_m - f_m\|_1. \tag{27}$$

The latter term can be bounded by using (Giné and Nickl, 2021, Proposition 4.3.33),

$$\|K_{h_N} * f_m - f_m\|_1 \le h_N^{[s]}\|f_m\|_{W_1^{[s]}(\mathbb{R}^2)}, \tag{28}$$

Note that there is a typo in (Giné and Nickl, 2021, Proposition 4.3.33) which states the result as in the one-dimensional case. The inequality (28) relies on the remainder term of a Taylor series. The remainder term in $\|\cdot\|_1$-norm is upper bounded by (Minkowski inequality for integrals)

$$[s]h_N^{[s]}\sum_{|\alpha|=[s]}\frac{1}{\alpha!}\int_0^1(1-u)^{[s]-1}\|D^\alpha f_m\|_1\,du \le [s]h_N^{[s]}\int_0^1(1-u)^{[s]-1}du\|f_m\|_{W_1^{[s]}(\mathbb{R}^2)},$$

where $\alpha = (\alpha_1, \alpha_2)$ is multi-index of dimension 2, $\alpha! = \alpha_1!\alpha_2!$, the integral $[s]\int_0^1 (1-u)^{[s]-1}$ is equal to 1, and $D^\alpha f_m$ is a weak-derivative of $f_m : \mathbb{R}^2 \to \mathbb{R}$. Hence,

$$N\|\widehat{f}^*_{m,N} - f_m\|_1 \leq N\|\widehat{f}^*_{m,N} - \mathbb{E}(\widehat{f}^*_{m,N})\|_1 + N\Lambda h_N^{[s]}$$

and with probability at least $1 - e^{-u}$,

$$N\|\widehat{f}^*_{m,N} - f_m\|_1 \leq N\Lambda h_N^{[s]} + (3/2)N\mathbb{E}\|\widehat{f}^*_{m,N} - \mathbb{E}(\widehat{f}^*_{m,N})\|_1 + \sqrt{2Nu}\|K\|_1 + u5\|K\|_1.$$

Recall that $\|f_m\|_{W_1^{[s]}(\mathbb{R}^2)} \leq \|f_m\|_{B^2_{1,\infty}(\mathbb{R}^2)}$ and, therefore, from (16) it follows that for any f_m such that $\|f_m\|_{B^2_{1,\infty}(\mathbb{R}^2)} \leq \mathit{\Lambda}$, with probability $1 - e^{-u}$ we have,

$$\|\widehat{f}^*_{m,N} - f_m\|_1 \leq \Lambda h_N^{[s]} + (3/2)\widetilde{C}N^{-\frac{[s]}{2[s]+2}} + \sqrt{2u/N}\|K\|_1 + u5\|K\|_1/N.$$

Substituting h_N as stated in Condition 2.ii, yields that with probability $1 - e^{-u}$,

$$\begin{aligned}&\|\widehat{f}^*_{m,N} - f_m\|_1\\ &\leq \left((3/2)\widetilde{C} + \mathit{\Lambda}(c\mathit{\Lambda})^{-\frac{[s]^2}{[s]+1}}\left(\frac{[s]-1}{2}\right)^{-\frac{[s]^2}{2[s]+2}}\right)N^{-\frac{[s]}{2[s]+2}} + \sqrt{2u/N}\|K\|_1 + u5\|K\|_1/N.\end{aligned}$$

Similarly, we can bound the difference between $\widehat{f}^*_{0,N}$ and f_0 in high probability; for $u > 0$, with probability $1 - e^{-u}$, it follows from (13) that

$$\begin{aligned}&\|f_0 \otimes f_0 - \widehat{f}^*_{0,N} \otimes \widehat{f}^*_{0,N}\|_1/(1+c_0) \leq \|\widehat{f}^*_{0,N} - f_0\|_1\\ &\leq \left((3/2)\widetilde{C} + \mathit{\Lambda}(c\mathit{\Lambda})^{-\frac{[s]^2}{[s]+1}}\left(\frac{[s]-1}{2}\right)^{-\frac{[s]^2}{2[s]+2}}\right)N^{-\frac{[s]}{2[s]+2}} + \sqrt{2u/N}\|K\|_1 + u5\|K\|_1/N.\end{aligned}$$

Substituting this into (14) gives that with probability at least $1 - 2e^{-u}$,

$$\begin{aligned}&|\beta(m) - \widehat{\beta}^*_N(m)|/(1+c_0/2)\\ &\leq \left((3/2)\widetilde{C} + \mathit{\Lambda}(c\mathit{\Lambda})^{-\frac{[s]^2}{[s]+1}}\left(\frac{[s]-1}{2}\right)^{-\frac{[s]^2}{2[s]+2}}\right)N^{-\frac{[s]}{2[s]+2}} + \sqrt{2u/N}\|K\|_1 + u5\|K\|_1/N.\end{aligned}$$

Furthermore, with probability

$$1 - N\beta(k^*) \geq 1 - 8C\log\Big(\frac{\gamma\eta}{8C} + \frac{3[s]+2}{2[s]+2}\log n\Big)n^{-\frac{[s]}{2[s]+2}}$$

$\widehat{f}^*_{0,N} = \widehat{f}_{0,N}$ and $\widehat{f}^*_{m,N} = \widehat{f}_{m,N}$. By setting $u = \log(n)[s]/(2[s]+2)$ we gain that with probability

$$1 - \Big(2 + 8C\log\Big(\frac{\gamma\eta}{8C} + \frac{3}{2}\log n\Big)\Big)n^{-\frac{[s]}{2[s]+2}}$$

the following bound holds

$$\begin{aligned}&|\beta(m) - \widehat{\beta}_N(m)|/(1+c_0/2)\\ &\leq \left(\frac{3}{2}\widetilde{C} + \mathit{\Lambda}(c\mathit{\Lambda})^{-\frac{[s]^2}{[s]+1}}\left(\frac{[s]-1}{2}\right)^{-\frac{[s]^2}{2[s]+2}}\right)N^{-\frac{[s]}{2[s]+2}} + \sqrt{\frac{[s]\log(n)}{([s]+1)N}}\|K\|_1 + \frac{5[s]\log(n)|K\|_1}{(2[s]+2)N}.\end{aligned}$$

Finally, observing that $N \geq (n/4k^*) - 1$, and noting that $16N^{-\frac{[s]}{2[s]+2}} \geq (N+1)^{-\frac{[s]}{2[s]+2}}$ whenever $N \geq 2$ and $[s] \geq 1$, and simplifying, we get,

$$\begin{aligned}
&|\beta(m) - \widehat{\beta}_N(m)| \\
&\leq 16\big(1 + \frac{c_0}{2}\big)\Bigg(\frac{3}{2}\widetilde{C} + \varLambda(c\varLambda)^{-\frac{[s]^2}{[s]+1}}\left(\frac{[s]-1}{2}\right)^{-\frac{[s]^2}{2[s]+2}} \\
&\qquad\qquad + \left(\sqrt{\frac{[s]}{([s]+1)}} + \frac{5[s]}{(2[s]+2)}\right)\log(n)|K\|_1\Bigg)\left(\frac{n}{4k^*}\right)^{-\frac{[s]}{2[s]+2}}.
\end{aligned}$$

Noting that $(4k^*)^{\frac{[s]}{2[s]+2}} \leq 4k^*$ and letting $C_1 = \frac{1}{\gamma}\log\left(\frac{\gamma\eta}{8C}\right)\left(\frac{3}{2}\widetilde{C} + \varLambda(c\varLambda)^{-\frac{[s]^2}{[s]+1}}\left(\frac{[s]-1}{2}\right)^{-\frac{[s]^2}{2[s]+2}}\right)$ and $C_2 = \frac{4\|K\|_1}{\gamma}\log\left(\frac{\gamma\eta}{8C}\right) + \frac{3}{2\gamma}\left(\frac{3}{2}\widetilde{C} + \varLambda(c\varLambda)^{-\frac{[s]^2}{[s]+1}}\left(\frac{[s]-1}{2}\right)^{-\frac{[s]^2}{2[s]+2}}\right)$, we obtain

$$|\beta(m) - \widehat{\beta}_N(m)| \leq 64\big(1 + \frac{c_0}{2}\big)(C_1 + C_2\log(n) + \frac{6\|K\|_1}{\gamma}\log^2(n))n^{-\frac{[s]}{2[s]+2}}.$$

□

Proof of Theorem 4. As in the proof of Theorem 3, we use a coupling argument together with concentration bounds on independent copies. More specifically, consider the geometrically ergodic Markov sample $X_0, \ldots, X_n$, and define the sequence of tuples $Z_i = (X_{2i(k+1)}, X_{2i(k+1)+m})$, $i = 0, 1, 2, \ldots, N-1$ where $N = N(k, n) := \lfloor \frac{n-k}{2(k+1)} \rfloor$ for some fixed $k \in m+1, \ldots, \lfloor n/8 \rfloor$; as in the continuous state-space setting, an optimal choice for k is specified later in the proof, see (35). As follows from Lemma 6 there exists a sequence of independent random vectors $Z_i^* = (Z_{i,1}^*, Z_{i,2}^*)$ for $i = 0, 1, \ldots, N-1$ each of which takes value in $\mathscr{X}^2$ and has the same distribution as Z_i such that

$$\Pr\left(\{\exists i \in 0, \ldots, N-1 : Z_i^* \neq Z_i\}\right) \leq N\beta(k) \tag{29}$$

Define

$$\widehat{\beta}_N^*(m) := \sum_{u\in\mathscr{X}}\sum_{v\in\mathscr{X}}|\widehat{P}_{m,N}^*(\{(u,v)\}) - \widehat{P}_{0,N}^*(\{u\})\widehat{P}_{0,N}^*(\{v\})|$$

where $\widehat{P}_{m,N}^*((u,v)) := \frac{1}{N}\sum_{i=0}^{N-1}\mathbf{1}_{\{(u,v)\}}(Z_i^*)$ and $\widehat{P}_{0,N}^*(u) := \frac{1}{N}\sum_{i=0}^{N}\mathbf{1}_{\{u\}}(Z_{i,1}^*)$. By a simple application of Jensen's inequality and noting that the random variables $Z_i^*,\ i = 0, \ldots, N-1$ are iid, for each $z \in \mathscr{X} \times \mathscr{X}$ we have,

$$\mathbb{E}[|\widehat{P}_{m,N}^*(z) - P_m(z)|] \leq \frac{1}{N}\Big(\sum_{i=0}^{N-1}\mathbb{E}(\mathbf{1}_{\{z\}}(Z_i^*) - \mathbb{E}\mathbf{1}_{\{z\}}(Z_i^*))^2\Big)^{1/2} \leq N^{-1/2} \tag{30}$$

where the second inequality is due to $\mathbb{V}\mathrm{ar}\big(\mathbf{1}_{\{z\}}(Z_i^*) - \mathbb{E}\mathbf{1}_{\{z\}}(Z_i^*)\big) \leq 1$. Similarly, for each $u \in \mathscr{X}$ we obtain $\mathbb{E}[|\widehat{P}_{0,N}^*(u) - P_0(u)|] \leq N^{-1/2}$. It follows that

$$\begin{aligned}
&\mathbb{E}|\widehat{\beta}_N^*(m) - \beta(m)| \\
&\leq \sum_{(u,v)\in\mathscr{X}^2}\mathbb{E}|P_m(\{(u,v)\}) - \widehat{P}_{m,N}^*(\{(u,v)\})| + 2\sum_{u\in\mathscr{X}}\mathbb{E}|P_0(\{u\}) - \widehat{P}_{0,N}^*(\{u\})| &(31)\\
&\leq 2|\mathscr{X}|^2 N^{-1/2}. &(32)
\end{aligned}$$

In much the same way as in the proof of Theorem 3, let $E := \{Z_i^* = Z_i,\ i \in 0, \dots, N-1\}$, and observe that $\mathbb{E}(|\widehat{\beta}_N^*(m) - \widehat{\beta}_N(m)||E) = 0$. Moreover, recall that $k \le \lfloor n/8 \rfloor$. We obtain,

$$\begin{aligned}\mathbb{E}|\beta(m) - \widehat{\beta}_N(m)| &\le \mathbb{E}|\beta(m) - \widehat{\beta}_N^*(m)| + 2N\beta(k) &(33)\\ &\le 2|\mathscr{X}|^2 \left(\frac{4k}{n-4k}\right)^{1/2} + 2n\eta e^{-\gamma k} \\ &\le 2|\mathscr{X}|^2 \left(\frac{8k}{n}\right)^{1/2} + 2n\eta e^{-\gamma k} \\ &\le \sqrt{32}|\mathscr{X}|^2 n^{-1/2} k + 2n\eta e^{-\gamma k} &(34)\end{aligned}$$

where (18) follows from (29). Optimizing (34) we obtain

$$k^\star = \frac{1}{\gamma} \log\left(\frac{\eta\gamma n^{3/2}}{\sqrt{8}|\mathscr{X}|^2}\right) \tag{35}$$

where n is taken large enough so ensure that $k^\star \ge 1$ (see (36) below). This choice of $k^\star$ and n leads to

$$\mathbb{E}|\widehat{\beta}_N(m) - \beta(m)| \le \frac{\sqrt{32}|\mathscr{X}|^2 n^{-1/2}}{\gamma} \left(1 + \log\left(\frac{\eta\gamma}{\sqrt{8}|\mathscr{X}|^2}\right) + \frac{3}{2}\log n\right).$$

Take $N = N(k^\star, n) = \lfloor \frac{n-k^\star}{2(k^\star+1)} \rfloor$, with $k^\star$ given by (35) and

$$n \ge \max\left\{2|\mathscr{X}|^3 \left(\frac{e^\gamma}{\eta\gamma}\right)^{2/3}, \left(\frac{\eta\gamma}{|\mathscr{X}|^2}\right)^{2/3}\right\}. \tag{36}$$

Substituting for $k^\star$ and noting that $N \le n/k^\star$ we have,

$$N\eta e^{-\gamma k^\star} \le \frac{\sqrt{2}|\mathscr{X}|^2 n^{-1/2}}{\log(\frac{\eta\gamma n^{3/2}}{\sqrt{8}|\mathscr{X}|^2})}. \tag{37}$$

On the other hand, by Hoeffding's inequality, for any $\epsilon > 0$ and each $u \in \mathscr{X}$ we have,

$$\Pr(|\widehat{P}_{0,N}^*(\{u\}) - P_0(\{u\})| \ge \frac{\epsilon}{2|\mathscr{X}|}) \le 2\exp\left\{-\frac{N\epsilon^2}{2|\mathscr{X}|^2}\right\} \tag{38}$$

Similarly, for each $(u,v) \in \mathscr{X}^2$ it holds that

$$\Pr(|\widehat{P}_{m,N}^*(\{(u,v)\}) - P_m(\{(u,v)\})| \ge \frac{\epsilon}{2|\mathscr{X}|^2}) \le 2\exp\left\{-\frac{N\epsilon^2}{2|\mathscr{X}|^4}\right\} \tag{39}$$

It follows that

$$\begin{aligned}&\Pr(|\widehat{\beta}_N^*(m) - \beta(m)| \ge \epsilon/2) \\ &\quad\le \sum_{u,v} \Pr(|P_m(\{(u,v)\}) - \widehat{P}_{m,N}^*(\{(u,v)\})| \ge \frac{\epsilon}{2|\mathscr{X}|^2}) \\ &\qquad\qquad + 2\sum_{u} \Pr(|P_0(\{v\}) - \widehat{P}_{0,N}^*(\{u\})| \ge \frac{\epsilon}{2|\mathscr{X}|}) \\ &\quad\le 2|\mathscr{X}|^2 \exp\left\{-\frac{N\epsilon^2}{2|\mathscr{X}|^4}\right\} + 2|\mathscr{X}| \exp\left\{-\frac{N\epsilon^2}{2|\mathscr{X}|^2}\right\}\end{aligned}$$

$$\leq 4|\mathscr{X}|^2 \exp\left\{-\frac{N\epsilon^2}{2|\mathscr{X}|^4}\right\}$$

$$\leq 4|\mathscr{X}|^2 \exp\{-\frac{(n-4k^\star)\epsilon^2}{8k^\star|\mathscr{X}|^4}\} \tag{40}$$

$$\leq 4|\mathscr{X}|^2 \exp\{-\frac{n\epsilon^2}{16k^\star|\mathscr{X}|^4}\} \tag{41}$$

$$\leq 4|\mathscr{X}|^2 \exp\{-\frac{\gamma n\epsilon^2}{16|\mathscr{X}|^4\left(\log\frac{\eta\gamma}{|\mathscr{X}|^2}+\frac{3}{2}\log_2 n\right)}\} \tag{42}$$

$$\leq 4|\mathscr{X}|^2 \exp\{-\frac{\gamma n\epsilon^2}{48|\mathscr{X}|^4\log n}\} \tag{43}$$

where, (40) follows from the choice of $N = \lfloor\frac{n-k^\star}{2(k^\star+1)}\rfloor$, (41) follows from recalling that in general, k (and thus also $k^\star$), is less than $\lfloor n/8\rfloor$, and finally, (42) and (43) follow from substituting the value of $k^\star$ as given by (24) and observing that by (36) we have $\frac{3}{2}\log n \geq \log(\frac{\eta\gamma}{|\mathscr{X}|^2})$. Hence, by (29), (37) and (43) we obtain,

$$\begin{aligned}
\Pr(|\widehat{\beta}_N(m)-\beta(m)|\geq\epsilon) &\leq N\beta(k^\star)+\Pr(|\widehat{\beta}^*_N(m)-\beta(m)|\geq\epsilon/2)\\
&\leq N\eta e^{-\gamma k^\star}+4|\mathscr{X}|^2\exp\left\{-\frac{N\epsilon^2}{2|\mathscr{X}|^4}\right\}\\
&\leq \frac{\sqrt{2}|\mathscr{X}|^2 n^{-1/2}}{\log(\frac{\eta\gamma n^{3/2}}{\sqrt{8}|\mathscr{X}|^2})}+4|\mathscr{X}|^2\exp\{-\frac{\gamma n\epsilon^2}{48|\mathscr{X}|^4\log n}\}
\end{aligned}$$

□

Proof of Theorem 5. As in the proof of Theorem 4, we start by a coupling argument, with the difference that instead of generating 2-tuples, we generate blocks of length $k+1$ for an appropriate value of k which we specify further in the proof. Specifically, given $X_0,\dots,X_n$ define $\widetilde{Z}_i = (X_{2i(k+1)}, X_{2i(k+1)+1},\dots,X_{(2i+1)k+2i})$, $i = 0,1,\dots,N-1$ where $N = N(k,n) := \lfloor\frac{n-k}{2(k+1)}\rfloor$ for some fixed $k \in 1,\dots,\lfloor n/8\rfloor$; an optimal choice for k is specified later in the proof, see (49). By Lemma 6 there exists a sequence of independent random vectors $\widetilde{Z}^*_i = (\widetilde{Z}^*_{i,0},\dots,\widetilde{Z}^*_{i,k})$ for $i = 0,1,\dots,N-1$ each of which takes value in $\mathscr{X}^{k+1}$ and has the same distribution as $\widetilde{Z}_i$ such that

$$\Pr\left(\{\exists i\in 0,\dots,N-1 : \widetilde{Z}^*_i \neq \widetilde{Z}_i\}\right) \leq N\beta(k). \tag{44}$$

Define

$$\widehat{\beta}^\dagger_N(m) := \sum_{u\in\mathscr{X}}\sum_{v\in\mathscr{X}} |\widehat{P}^\dagger_{m,N}(\{(u,v)\}) - \widehat{P}^\dagger_{0,N}(\{u\})\widehat{P}^\dagger_{0,N}(\{v\})|$$

where $\widehat{P}^\dagger_{m,N}((u,v)) := \frac{1}{N}\sum_{i=0}^{N-1}\mathbf{1}_{\{(u,v)\}}(\widetilde{Z}^*_{i,0},\widetilde{Z}^*_{i,m})$ and $\widehat{P}^\dagger_{0,N}(u) := \frac{1}{N}\sum_{i=0}^{N}\mathbf{1}_{\{u\}}(\widetilde{Z}^*_{i,0})$. As in the proof of Theorem 4, for each $u\in\mathscr{X}$ it holds that

$$\mathbb{E}[|\widehat{P}^\dagger_{0,N}(u) - P_0(u)|] \leq N^{-1/2}. \tag{45}$$

Define the class of indicator functions

$$\mathcal{H}_k = \{h_{m,z} : \mathscr{X}^k \to \{0,1\} : z\in\mathscr{X}\times\mathscr{X},\ h_{m,z}(\mathbf{x}) := \mathbf{1}_{\{z\}}(x_0,x_m),\ m=1,\dots,k\}.$$

It is straightforward to verify that the VC-dimension of $\mathcal{H}_m$ is at most $\log_2(|\mathscr{X}|k)$. Therefore, as follows from (Giné and Nickl, 2021, pp. 217) it holds that,

$$\mathbb{E}[\sup_{\substack{m\in 1,\dots,k\\ (u,v)\in\mathscr{X}^2}} |P^\dagger_{m,N}((u,v)) - P_m((u,v))|] \leq \sqrt{\frac{8\log_2(|\mathscr{X}|k)\log N}{N}} \tag{46}$$

By (45) and (46) we have,

$$\begin{aligned}
&\mathbb{E}[\sup_{m\in 1,\dots,k} |\widehat{\beta}^{\dagger}_N(m) - \beta(m)|] \\
&\leq \mathbb{E}[\sup_{m\in 1,\dots k} \sum_{(u,v)\in\mathscr{X}^2} |P_m(\{(u,v)\}) - \widehat{P}^{\dagger}_{m,N}(\{(u,v)\})| + 2\sum_{u\in\mathscr{X}} |P_0(\{u\}) - \widehat{P}^{\dagger}_{0,N}(\{u\})|] \\
&\leq |\mathscr{X}|^2 \mathbb{E}[\sup_{\substack{m\in 1,\dots k\\ (u,v)\in\mathscr{X}^2}} |P_m(\{(u,v)\}) - \widehat{P}^{\dagger}_{m,N}(\{(u,v)\})|] + 2\sum_{u\in\mathscr{X}} \mathbb{E}|P_0(\{u\}) - \widehat{P}^{\dagger}_{0,N}(\{u\})| \\
&\leq |\mathscr{X}|^2 \sqrt{\frac{8\log_2(|\mathscr{X}|k)\log N}{N}} + 2|\mathscr{X}|N^{-1/2} \\
&\leq 2|\mathscr{X}|^2 \sqrt{\frac{8\log_2(|\mathscr{X}|k)\log N}{N}}
\end{aligned} \tag{47}$$

Together with the coupling argument given earlier we obtain,

$$\begin{aligned}
\mathbb{E}[\sup_{m\in 1,\dots,k} |\widehat{\beta}_N(m) - \beta(m)|] &\leq \mathbb{E}[\sup_{m\in 1,\dots,k} |\widehat{\beta}^{\dagger}_N(m) - \beta(m)|] + N\beta(k) \\
&\leq 2|\mathscr{X}|^2 \sqrt{\frac{8\log_2(|\mathscr{X}|k)\log N}{N}} + N\eta e^{-\gamma k} \\
&\leq 2|\mathscr{X}|^2 \log(N|\mathscr{X}|)\sqrt{8} N^{-1/2} + N\eta e^{-\gamma k} \\
&\leq 8|\mathscr{X}|^2 \sqrt{\frac{2k}{n}} \log(n|\mathscr{X}|) + n\eta e^{-\gamma k} \\
&\leq 8\sqrt{2}|\mathscr{X}|^2 n^{-1/2} \log(n|\mathscr{X}|)k + n\eta e^{-\gamma k}
\end{aligned} \tag{48}$$

Optimizing (48) we have

$$k^{\dagger} = \frac{1}{\gamma}\log\left(\frac{\eta\gamma n^{3/2}}{8\sqrt{2}|\mathscr{X}|^2\log(n|\mathscr{X}|)}\right) \tag{49}$$

with $n \geq \frac{8\sqrt{2}|\mathscr{X}|^3 e^{\gamma}}{\eta\gamma}$ to ensure that $k^{\dagger} \geq 1$. This leads to

$$\begin{aligned}
&\mathbb{E}[\sup_{m\in 1,\dots,k} |\widehat{\beta}_N(m) - \beta(m)|] \\
&\qquad \leq \frac{8\sqrt{2}|\mathscr{X}|^2 n^{-1/2}\log(n|\mathscr{X}|)}{\gamma}\left(1 + \log\left(\frac{\eta\gamma n^{1/2}}{8\sqrt{2}|\mathscr{X}|^2\log(n|\mathscr{X}|)}\right)\right).
\end{aligned} \tag{50}$$

Take $N = N(k^{\dagger}, n) = \lfloor \frac{n-k^{\dagger}}{2(k^{\dagger}+1)} \rfloor$, with $k^{\dagger}$ given by (49) and

$$n \geq \max\left\{\frac{8\sqrt{2}|\mathscr{X}|^3 e^{\gamma}}{\eta\gamma}, \frac{\eta\gamma}{8\sqrt{2}|\mathscr{X}|}\right\}. \tag{51}$$

It follows that,

$$N\eta e^{-\gamma k^{\dagger}} \leq \frac{4\sqrt{2}|\mathscr{X}|^2\log(n|\mathscr{X}|)n^{-1/2}}{\log(\frac{\eta\gamma n^{3/2}}{8\sqrt{2}|\mathscr{X}|^2})} \tag{52}$$

On the other hand, by Hoeffding's inequality, for any $\epsilon > 0$ and $u \in \mathscr{X}$ it holds that

$$\Pr(|\widehat{P}^{\dagger}_{0,N}(\{u\}) - P_0(\{u\})| \geq \frac{\epsilon}{2|\mathscr{X}|}) \leq 2\exp\left\{-\frac{N\epsilon^2}{2|\mathscr{X}|^2}\right\} \tag{53}$$

Furthermore, noting that $\mathcal{H}_m$ is a VC-class, by (Devroye et al., 2013, Theorem 12.5) for $\epsilon > 0$ we have,

$$\Pr\Big(\sup_{\substack{m\in 1,\dots,k^\dagger\\(u,v)\in\mathscr{X}^2}} |\widehat{P}^\dagger_{m,N}(\{(u,v)\}) - P_m(\{(u,v)\})| \geq \frac{\epsilon}{2|\mathscr{X}|^2}\Big) \leq 8\log_2(|\mathscr{X}|k)e^{-\frac{N\epsilon^2}{128|\mathscr{X}|^4}} \tag{54}$$

Therefore, for any $\epsilon > 0$ we have,

$$\begin{aligned}
&\Pr\Big(\sup_{m\in 1,\dots,k^\dagger} |\widehat{\beta}^\dagger_N(m) - \beta(m)| \geq \epsilon/2\Big)\\
&\qquad\leq |\mathscr{X}|^2 \Pr\left(\sup_{\substack{m\in 1,\dots,k^\dagger\\(u,v)\in\mathscr{X}^2}} |P_m(\{(u,v)\}) - \widehat{P}^\dagger_{m,N}(\{(u,v)\})| \geq \frac{\epsilon}{2|\mathscr{X}|^2}\right)\\
&\qquad\qquad + 2\sum_u \Pr\left(|P_0(\{v\}) - \widehat{P}^\dagger_{0,N}(\{u\})| \geq \frac{\epsilon}{2|\mathscr{X}|}\right)\\
&\qquad\leq 8|\mathscr{X}|^2\log(|\mathscr{X}|k^\dagger)\exp\left\{-\frac{N\epsilon^2}{128|\mathscr{X}|^4}\right\} + 2|\mathscr{X}|\exp\left\{-\frac{N\epsilon^2}{2|\mathscr{X}|^2}\right\} &(55)\\
&\qquad\leq 16|\mathscr{X}|^2\log(|\mathscr{X}|k^\dagger)\exp\left\{-\frac{N\epsilon^2}{128|\mathscr{X}|^4}\right\}\\
&\qquad\leq 16|\mathscr{X}|^2\log(|\mathscr{X}|k^\dagger)\exp\{-\frac{n\epsilon^2}{1024k^\dagger|\mathscr{X}|^4}\} &(56)\\
&\qquad\leq 16|\mathscr{X}|^2\log\left(\frac{3|\mathscr{X}|}{2\gamma}\log(\eta^{2/3}\gamma^{2/3}n)\right)\exp\left\{-\frac{\gamma n\epsilon^2}{3072|\mathscr{X}|^4\log n}\right\} &(57)
\end{aligned}$$

where, (55) follows from (53) and (54), (56) follows from the choice of $N = \lfloor\frac{n-k^\dagger}{2(k^\dagger+1)}\rfloor$ and noting that in general all k (and thus also $k^\dagger$) are taken to be less than $\lfloor n/8\rfloor$, (57) follows from substituting for the value of $k^\dagger$ as given by (49) and observing that by (51) we have $\frac{n^{3/2}}{\log(n|\mathscr{X}|)} \geq \frac{n}{|\mathscr{X}|} \geq \frac{\eta\gamma}{8\sqrt{2}|\mathscr{X}|^2}$. Hence, by (44), (52) and (57) we obtain,

$$\begin{aligned}
&\Pr\Big(\sup_{m=1,\dots,k^\dagger} |\widehat{\beta}_N(m) - \beta(m)| \geq \epsilon\Big)\\
&\quad\leq N\eta e^{-\gamma k^\dagger} + \Pr\Big(\sup_{m=1,\dots,k^\dagger} |\widehat{\beta}^\dagger_N(m) - \beta(m)| \geq \epsilon/2\Big)\\
&\quad\leq \frac{4\sqrt{2}|\mathscr{X}|^2\log(n|\mathscr{X}|)n^{-1/2}}{\log(\frac{\eta\gamma n^{3/2}}{8\sqrt{2}|\mathscr{X}|^2})}\\
&\qquad\qquad + 16|\mathscr{X}|^2\log\left(\frac{3|\mathscr{X}|}{2\gamma}\log(\eta^{2/3}\gamma^{2/3}n)\right)\exp\left\{-\frac{\gamma n\epsilon^2}{3072|\mathscr{X}|^4\log n}\right\}
\end{aligned}$$

and the result follows. □

References

M. A. Arcones and B. Yu. Central limit theorems for empirical and u-processes of stationary mixing sequences. *Journal of Theoretical Probability*, 7:47–71, 1994.

C. Bennett and R. C. Sharpley. *Interpolation of Operators*. Elsevier Science, 1988.

H. C. Berbee. Random walks with stationary increments and renewal theory. *MC Tracts*, 1979.

P. Bertail, P. Doukhan, and P. Soulier. *Dependence in probability and statistics*. Springer, 2006.

D. Bosq. *Nonparametric statistics for stochastic processes: estimation and prediction*, volume 110. Springer Science & Business Media, 2012.

R. C. Bradley. *Introduction to Strong Mixing Conditions*, volume 1,2, 3. Kendrick Press, 2007.

J. Dedecker and C. Prieur. New dependence coefficients. examples and applications to statistics. *Probability Theory and Related Fields*, 132:203–236, 2005.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2021.

D. Hsu., A. Kontorovich, D. A. Levin, Y. Peres, C. Szepesvári, and G. Wolfer. Mixing time estimation in reversible Markov chains from a single sample path. *The Annals of Applied Probability*, 29(4):2439 – 2480, 2019.

I. A. Ibragimov. Some limit theorems for stationary processes. *Theory of Probability & its Applications*, 7 (4):349–382, 1962.

A. Khaleghi and G. Lugosi. Inferring the mixing properties of a stationary ergodic process from a single sample-path. *IEEE Transactions on Information Theory*, 69(6):4014–4026, 2023.

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

S. Levental. Uniform limit theorems for harris recurrent markov chains. *Probability theory and related fields*, 80(1):101–118, 1988.

D. McDonald, C. Shalizi, and M. Schervish. Estimating beta-mixing coefficients via histograms. *Electronic Journal of Statistics*, 9(2):2855–2883, 2015.

A. B. Nobel. Hypothesis testing for families of ergodic processes. *Bernoulli*, 12(2):251–269, 2006. ISSN 13507265.

E. Rio. *Théorie asymptotique des processus aléatoires faiblement dépendants*. Springer, 1999.

E. Rio. Inégalités de hoeffding pour les fonctions lipschitziennes de suites dépendantes. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 330(10):905–908, 2000.

P. M. Samson. Concentration of measure inequalities for markov chains and ϕ-mixing processes. *The Annals of Probability*, 2000.

P. Shields. *The Ergodic Theory of Discrete Sample Paths*. AMS Bookstore, 1996.

G. Viennet. Inequalities for absolutely regular sequences: application to density estimation. *Probability theory and related fields*, 107:467–492, 1997.

G. Wolfer. Mixing time estimation in ergodic markov chains from a single trajectory with contraction methods. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 890–905, 2020.

G. Wolfer and A. Kontorovich. Estimating the mixing time of ergodic markov chains. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 3120–3159, 25–28 Jun 2019.

B. Yu. Rates of Convergence for Empirical Processes of Stationary Mixing Sequences. *The Annals of Probability*, 22(1):94 – 116, 1994.