Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/235118/

Version: Accepted Version

Estimating the Mixing Coefficients of Geometrically Ergodic Markov Processes

Steffen Grünewälder*, Azadeh Khaleghi†

Abstract—We propose methods to estimate the β-mixing coefficients of a real-valued geometrically ergodic Markov process from a single sample-path $X_0, X_1, \ldots, X_{n-1}$. 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 $\mathcal{O}(\log(n) n^{-1/2})$. The theoretical results are complemented with empirical evaluations.

Index Terms—β-mixing, estimation, geometric ergodicity, Markov process, consistency

I. 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. [1]–[9] 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.

* Department of Mathematics, University of York, York, UK, steffen.grunewalder@york.ac.uk
† ENSAE - CREST, Institut Polytechnique de Paris, Palaiseau, France, azadeh.khaleghi@ensae.fr
[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 $\{s\} \in (0, 1]$.

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. A stationary, geometrically ergodic Markov process is absolutely regular with $\beta(m) \to 0$ *at least exponentially fast* as $m \to 0$. We start by recalling the relevant concepts.

β-mixing coefficients. Let $(\Omega, \mathcal{F}, \mu)$ be a probability space. The β-dependence $\beta(\mathcal{U}, \mathcal{V})$ between the σ-subalgebras $\mathcal{U}$ and $\mathcal{V}$ of $\mathcal{F}$ is defined as follows. Let $\iota(\omega) \mapsto (\omega, \omega)$ be the injection map from $(\Omega, \mathcal{F})$ 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 $(\Omega \times \Omega, \mathcal{U} \otimes \mathcal{V})$ 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

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

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

$$\beta(m) := \sup_{j \in \mathbb{N}} \beta(\sigma(\{X_t : 0 \leq t \leq j-1\}), \sigma(\{X_t : t \geq j+m\})).$$

A stochastic process is said to be β-mixing or absolutely regular if $\lim_{m \to \infty} \beta(m) = 0$.

Geometrically ergodic Markov processes. 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 $(X_\ell, \ldots, X_{m+\ell})$ is the same as that of $(X_0, \ldots, X_m)$. 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_0 and X_m respectively, i.e.

$$\beta(m) = \beta(\sigma(X_0), \sigma(X_m))$$

[7, 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

$$\sup_{B \in \mathcal{B}(\mathbb{R})} |p_m(x, B) - \rho(B)| \leq f(x) e^{-c(x)m}$$

where $p_m(x, B)$ defined for $x \in \mathbb{R}$ and $B \in \mathcal{B}(\mathbb{R})$ is the regular conditional probability distribution of X_m given X_0 and ρ denotes the marginal distribution of X_0 [7, vol. 2

Definition 21.18 pp 325]. It is well-known - see, e.g. [7, 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 β-mixing coefficients of the process satisfy

$$\beta(m) \leq \eta e^{-\gamma m} \tag{2}$$

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 $m = 1, 2, \ldots$, of a real-valued geometrically ergodic Markov process from a finite sample path $X_0, X_1, \ldots, X_n$. 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 $B_{1,\infty}^s(\mathbb{R}^2)$; roughly speaking, this implies that f_m has $[s]$ many weak derivatives. As discussed above, for a geometrically ergodic Markov process, there exists some $\eta, \gamma > 0$ such that $\beta(m) \leq \eta e^{-\gamma m}$. Given some $\eta, \ \gamma \in (0, \infty)$ we show in Theorem 3 that

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

for all $m \ \lesssim \ (\log n)/\gamma$, where, $\widehat{\beta}_N$ is given by (7) 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 5 and 6 are concerned with a different setting where the state-space of the Markov processes 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 (10) with $N \approx (\log n)/\gamma$ and every $m \lesssim (\log n)/\gamma$ we have,

$$\mathbb{E}|\widehat{\beta}_N(m) - \beta(m)| \lesssim \frac{|\mathcal{X}|}{\gamma} \log(n) n^{-1/2}. \tag{3}$$

Moreover,

$$\Pr(|\widehat{\beta}_N(m) - \beta(m)| \geq \epsilon)$$
$$\lesssim |\mathcal{X}| n^{-1/2} + |\mathcal{X}|^2 \exp\left(-\frac{\gamma n \epsilon^2}{|\mathcal{X}|^4 \log n}\right)$$

for $\epsilon > 0$. We refer to the statement of Theorem 5 for the explicit constants. Observe that we have a factor $|\mathcal{X}|/\gamma$ in (3). In other words, γ has the same effect here as in the bound provided in Theorem 3, and the constant increases linearly 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 6 for a specification). The analysis relies

on a VC-argument in place of a union bound which leads to tighter error bounds in this case. We obtain,

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

and,

$$\Pr(\sup_{m \leq k^\dagger} |\widehat{\beta}_N(m) - \beta(m)| \geq \epsilon)$$
$$\lesssim |\mathcal{X}|^2 \log(n|\mathcal{X}|) n^{-1/2} + |\mathcal{X}|^2 \exp\left(-\frac{\gamma n \epsilon^2}{|\mathcal{X}|^4 \log n}\right).$$

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.

In our empirical evaluations provided in Section V, we highlight one of the key features of our estimators: their model-agnostic nature, which enables them to function effectively without prior knowledge of the underlying time-series models, requiring only assumptions on the smoothness of densities and loose bounds on the mixing rate. We specifically demonstrate that a natural model-specific β-mixing estimator, designed for *autoregressive AR* process, suffers performance drop in the presence of model-mismatch. In contrast to our estimator, this model-specific estimator exhibits bias and fails to consistently estimate the mixing coefficient, highlighting its vulnerability to inaccuracies when the underlying model is not correctly specified.

Related literature. Research on the direct estimation of the mixing coefficients is relatively scarce. In the asymptotic regime, [10] used hypothesis testing to give asymptotically consistent estimates of the polynomial decay rates for covariance-based mixing conditions. [11] 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. [12]), 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 [13]. 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 [14, pp. 404], it is highly unlikely that a dimension-independent rate would be achievable for an estimator of the β-mixing coefficient. An interesting body of work exists for a different, yet related problem, concerning the estimation of the mixing times of finite-state Markov processes [15]–[17]. Indeed, in their recent work, [18] strategically exploit the relationship between the mixing time and the β-mixing coefficients of an ergodic Markov process with a discrete alphabet, and use this connection to construct estimators for the average mixing time of the process. Moreover,

they propose an approach that reduces the dependence on the size of the state space in the discrete-alphabet setting. This in turn, enables them to handle certain countably infinite state-space chains and to estimate the corresponding mixing coefficients.

Organization. The remainder of the paper is organized as follows. In Section II we introduce some preliminary notation and basic definitions. After formulating the problem in Section III, we propose our estimators in Section IV, considering real-valued as well as we finite-valued sample-paths; the technical proofs are deferred to Section VII. In Section V, we present empirical evaluations of our estimators, focusing on their model-agnostic nature, to support the primary theoretical results. In Section VI we provide concluding remarks and explore potential future research directions.

II. 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 $\mathcal{U}$ and $\mathcal{V}$ are two σ-algebras, we will occasionally use standard measure-theoretic notation as follows: their product σ-algebra is denoted by $\mathcal{U} \otimes \mathcal{V} := \sigma(\mathcal{U} \times \mathcal{V})$, and the σ-algebra generated by their union is denoted by $\mathcal{U} \vee \mathcal{V} := \sigma(\mathcal{U} \cup \mathcal{V})$. 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 IV-A, we impose classical density assumptions on certain finite-dimensional marginals of the Markov processes considered. The densities satisfy standard smoothness conditions as controlled by the parameters of appropriate Besov spaces.

Besov Spaces $B_{p,\infty}^s(\mathbb{R}^d)$. 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 $\Delta_h^r f := (\Delta_h \circ \Delta_h^{r-1}) f$ 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 $B_{p,\infty}^s(\mathbb{R}^d)$ is defined as

$$B_{p,\infty}^s(\mathbb{R}^d) := \{f \in L_p(\mathbb{R}^d) : \|f\|_{B_{p,\infty}^s(\mathbb{R}^d)} < \infty\} \quad (4)$$

with Besov norm

$$\|f\|_{B_{p,\infty}^s(\mathbb{R}^d)} := \|f\|_1 + \sup_{0 < t < \infty} t^{-s} \sup_{|h| \leq t} \|\Delta_h^r f\|_1$$

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 $r > s$ [19]. 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 $B_{p,\infty}^s(\mathbb{R}^d)$.

Remark 1. *As follows from [19, 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*

$$W_p^{r_0}(\mathbb{R}^d) \cap W_p^{r_1}(\mathbb{R}^d) \hookrightarrow B_{p,\infty}^\varsigma(\mathbb{R}^d) \hookrightarrow W_p^{r_0}(\mathbb{R}^d) + W_p^{r_1}(\mathbb{R}^d). \quad (5)$$

If $r_0 < r_1$, then the left and right hand sides of (5) reduce to $W_p^{r_1}(\mathbb{R}^d)$ and $W_p^{r_0}(\mathbb{R}^d)$ respectively. Furthermore, in this case we have $\|f\|_{B_{p,\infty}^\varsigma(\mathbb{R}^d)} \geq \|f\|_{W_p^{r_0}(\mathbb{R}^d)}$. 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 [19, Theorem 5.4.14 and Definition 5.1.7] we have

$$\|f\|_{B_{p,\infty}^\varsigma(\mathbb{R}^d)}$$
$$= \sup_{t>0} t^{-\theta} K(f, t; W_p^{r_0}(\mathbb{R}^d), W_p^{r_1}(\mathbb{R}^d))$$
$$\geq K(f, 1; W_p^{r_0}(\mathbb{R}^d), W_p^{r_1}(\mathbb{R}^d))$$
$$\geq \inf\{\|f_0\|_{W_p^{r_0}(\mathbb{R}^d)} + \|f_1\|_{W_p^{r_0}(\mathbb{R}^d)} : f = f_0 + f_1\}$$
$$\geq \|f\|_{W_p^{r_0}(\mathbb{R}^d)}$$

In particular, consider the Besov space $B_{1,\infty}^s(\mathbb{R}^d)$ 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,

$$\|f\|_{B_{1,\infty}^s(\mathbb{R}^d)} \geq \|f\|_{W_1^{[s]}(\mathbb{R}^d)}.$$

III. PROBLEM FORMULATION

We are given a sample $X_0, X_1, \dots, X_n$ generated by a stationary geometrically ergodic Markov process taking values in some $\mathcal{X} \subseteq \mathbb{R}$. As discussed in Section II such a process is known to have a sequence of β-mixing coefficients such that

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

for some unknown constants η, $\gamma \in (0, \infty)$. The mixing coefficient $\beta(m)$ is unknown, and it is our objective to estimate it. In Section IV we introduce our estimators together with theoretical bounds on the estimation error. In Section V we demonstrate the performance of the proposed estimators on samples generated by standard time-series models.

IV. THEORETICAL RESULTS

In this section, we present our estimators for $\beta(m)$. We consider two different settings, depending on the state-space. First, in Section IV-A, 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 IV-B, we consider the case where the state-space $\mathcal{X}$ 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 6. 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 $U \in \mathcal{B}(\mathcal{X}^2)$ we have

$$\Pr(\{(X_0, X_m) \in U\}) = P_m(U).$$

By stationarity,

$$\Pr(\{(X_t, X_{m+t}) \in U\}) = P_m(U), \ t \in \mathbb{N}.$$

A. Real-valued state-space

We start by considering the case where $\mathcal{X}$ 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 $k \in m, m+1, \dots, \lfloor n/8 \rfloor$ let $N = N(k,n) := \lfloor \frac{n-k}{2(k+1)} \rfloor$ and define the sequence of tuples

$$Z_i = (X_{2i(k+1)}, X_{2i(k+1)+m}), \ i = 0, 1, 2, \dots, N.$$

Define the Kernel Density Estimator (KDE) of f_m as

$$\widehat{f}_{m,N}(z) = \frac{1}{Nh_N^2}\sum_{i=0}^{N-1} K\left(\frac{z - Z_i}{h_N}\right) \tag{6}$$

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

$$\widehat{\beta}_N(m) = \frac{1}{2}\int_{\mathbb{R}^2} |\widehat{f}_{m,N} - \widehat{f}_{0,N} \otimes \widehat{f}_{0,N}| d\lambda_2 \tag{7}$$

where $\otimes$ denotes the tensor product. Note that to simplify notation in (6) and (7), we have omitted the dependence of N on the choice k. An optimal value for k, denoted by $k^\star$, 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 (6). Recall that a bivariate kernel is said to be of order ℓ if

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

$$\int_{z\in\mathbb{R}^2} K(z)d\lambda_2(z) = 1$$

and for all $i, \ j \in \mathbb{N}, \ i+j < \ell$,

$$\int_{z=(z_1,z_2)\in\mathbb{R}^2} z_1^i z_2^j K(z)d\lambda_2(z) = 0.$$

Condition 2. *The density $f_m \in B_{1,\infty}^s(\mathbb{R}^2)$ for some $s > 1$ and $\|f_m\|_{B_{1,\infty}^s(\mathbb{R}^2)} \le \Lambda$ for some $\Lambda \in (0,\infty)$. It is further assumed that the ℓth moments of the pair (X_0, X_m) are finite for $\ell = 1, \dots, \lceil s \rceil$, and*

$$\int_{\mathbb{R}^2} f_m(z)(1 + \|z\|^2)d\lambda_2(z) < \infty$$

$$\int_{\mathbb{R}^2} K^2(z)(1 + \|z\|^2)d\lambda_2(z) < \infty$$

We assume that some $(\eta, \gamma) \in (0,\infty)^2$ is provided satisfying $\beta(m) \le \eta e^{-\gamma m}$. The estimator $\widehat{\beta}_N(m)$ given by (7) for $m = 1, \dots, k^\star$ 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 $N = \lfloor \frac{n-k^\star}{2(k^\star+1)} \rfloor$ with

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

and

$$C := (2 + c_0)(L_1)^{\frac{[s]}{[s]+1}}(c\Lambda)^{\frac{1}{[s]+1}}$$

with L_1 defined such that L_1^2 is given by,

$$\sqrt{2}\int_{\mathbb{R}^2} f(z)(1 + \|z\|^2)d\lambda_2(z)\int_{\mathbb{R}^2} K^2(z)(1 + \|z\|^2)d\lambda_2(z)$$

by choosing the hyperparameter in [14, Remark 5.16] to be equal to 2.

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 (7), 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. Note that the corresponding Besov norm will be influenced by σ. In particular, for very small values of σ the densities will have a sharp peak and the Besov norm will be large.

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

$$\mathbb{E}|\beta(m) - \widehat{\beta}_N(m)|$$

$$\le \frac{8Cn^{-\frac{[s]}{2[s]+2}}}{\gamma}\left(e^\gamma + \log\frac{\gamma\eta}{8C} + \left(\frac{3[s]+2}{2[s]+2}\right)\log n\right)$$

Moreover, with probability at least $1 - 8Cn^{-\frac{[s]}{2[s]+2}}$ it holds that

$$|\beta(m) - \widehat{\beta}_N(m)|$$

$$\le 64\left(1 + \frac{c_0}{2}\right)(C_1 + C_2\log(n) + \frac{6\|K\|_1}{\gamma}\log^2(n))n^{-\frac{[s]}{2[s]+2}}.$$

where

$$C_1 := \frac{1}{\gamma}\log\left(\frac{\gamma\eta}{8C}\right)(A_1 + B_1),$$

with

$$A_1 := 3(L_1)^{\frac{[s]}{[s]+1}}(c\Lambda)^{\frac{1}{[s]+1}}$$

$$B_1 := \Lambda(c\Lambda)^{-\frac{[s]^2}{[s]+1}}\left(\frac{[s]-1}{2}\right)^{-\frac{[s]^2}{2[s]+2}}$$

and,

$$C_2 := \frac{4\|K\|_1}{\gamma} \log\left(\frac{\gamma\eta}{8C}\right) + \frac{3}{2\gamma}(A_2 + B_2)^{-\frac{[s]^2}{2[s]+2}}$$

where

$$A_2 := 3(L_1)^{\frac{[s]}{[s]+1}}(c\Lambda)^{\frac{1}{[s]+1}}$$

$$B_2 := \Lambda(c\Lambda)^{-\frac{[s]^2}{[s]+1}}\left(\frac{[s]-1}{2}\right).$$

See Section VII for a proof.

Remark 4. *We refer to Section V for a discussion on natural choices for $k^\star$, η and γ in the context of some standard time-series models such as AR(1).*

B. Finite state-space

In the special case where the state-space $\mathcal{X}$ of the Markov process 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 $\{X_t = s\}, s \in \mathcal{X}$. Therefore, by [7, vol. Proposition 3.21 pp. 88] we have

$$\beta(m) = \frac{1}{2}\sum_{u\in\mathcal{X}}\sum_{v\in\mathcal{X}}|P_m(\{(u,v)\}) - P_0(\{u\})P_0(\{v\})| \quad (9)$$

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 (9) as follows. Fix a lag of length $k \in 1, \ldots, n$ (an optimal value for which will be specified in Proposition 5). Define the sequence of tuples $Z_i = (X_{2ki}, X_{2k(i+1)})$, $i = 0, 2, 4, \ldots, N-1$, with $N = N(k, n) := \lfloor\frac{n-k}{2(k+1)}\rfloor$. For each pair $(u, v) \in \mathcal{X}^2$ let

$$\widehat{P}_{m,N}((u,v)) = \frac{1}{N}\sum_{i=0}^{N-1}\mathbf{1}_{\{(u,v)\}}(Z_i).$$

Similarly, for each $u \in \mathcal{X}$ we can obtain an empirical estimate of $P_0(u)$ as

$$\widehat{P}_{0,N}(u) := \frac{1}{2N}\sum_{i=0}^{2N-1}\mathbf{1}_{\{u\}}(X_{ki}).$$

Define $\widehat{\beta}_N(m)$ as

$$\frac{1}{2}\sum_{u\in\mathcal{X}}\sum_{v\in\mathcal{X}}|\widehat{P}_{m,N}(\{(u,v)\}) - \widehat{P}_{0,N}(\{u\})\widehat{P}_{0,N}(\{v\})| \quad (10)$$

Theorem 5. *Consider a sample of length $n \in \mathbb{N}$ of a stationary, geometrically ergodic Markov process with finite state-space $\mathcal{X}$. Define $N(k, n) = \lfloor\frac{n-k}{2(k+1)}\rfloor$, $n, k \in \mathbb{N}$. Let $k^\star := \lfloor\frac{1}{\gamma}\log\left(\frac{\eta\gamma n^{3/2}}{\sqrt{2}|\mathcal{X}|}\right)\rfloor$ and $N = N(k^\star, n)$ with*

$$n \geq \max\left\{\left(\frac{\sqrt{2}|\mathcal{X}|e^{2\gamma}}{\eta\gamma}\right)^{2/3}, \left(\frac{\eta\gamma}{|\mathcal{X}|}\right)^{2/3}\right\}.$$

For every $m = 1, \ldots, k^\star$ and $\widehat{\beta}_N(m)$ given by (10) we have,

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

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

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

The proof is provided in Section VII.

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 6. 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 $\log k^\dagger$ factor.

Theorem 6. *Consider a sample of length $n \in \mathbb{N}$ of a stationary, geometrically ergodic Markov process with finite state-space $\mathcal{X}$. For n, $k \in \mathbb{N}$ define*

$$N(k, n) = \left\lfloor\frac{n-k}{2(k+1)}\right\rfloor.$$

Let $N = N(k^\dagger, n)$ with

$$k^\dagger := \left\lfloor\frac{1}{\gamma}\log\left(\frac{\eta\gamma n^{3/2}}{4\sqrt{2}|\mathcal{X}|^2\log_2(n|\mathcal{X}|)}\right)\right\rfloor$$

and

$$n \geq \max\left\{\frac{4\sqrt{2}|\mathcal{X}|^3e^\gamma}{\eta\gamma}, \frac{2k(k+1)}{3k+2}\right\}.$$

For $\widehat{\beta}_N$ given by (10) we have,

$$\mathbb{E}[\sup_{m\in 1,\ldots,k^\dagger}|\widehat{\beta}_N(m) - \beta(m)|]$$
$$\leq \frac{4\sqrt{2}|\mathcal{X}|^2n^{-1/2}\log_2(n|\mathcal{X}|)}{\gamma}$$
$$\times\left(e + \log\left(\frac{\eta\gamma n^{3/2}}{4\sqrt{2}|\mathcal{X}|^2\log_2(n|\mathcal{X}|)}\right)\right)$$

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

$$\Pr(\sup_{m\in 1,\ldots,k^\dagger}|\widehat{\beta}_N(m) - \beta(m)| \geq \epsilon)$$
$$\leq \frac{4\sqrt{2}|\mathcal{X}|^2\log(n|\mathcal{X}|)n^{-1/2}}{\log(\frac{\eta\gamma n^{3/2}}{8\sqrt{2}|\mathcal{X}|^2})}$$
$$+ 16|\mathcal{X}|^2\log\left(\frac{3|\mathcal{X}|}{2\gamma}\log(\eta^{2/3}\gamma^{2/3}n)\right)$$
$$\times\exp\left\{-\frac{\gamma n\epsilon^2}{3072|\mathcal{X}|^4\log n}\right\}$$

The proof is provided in Section VII.

V. EMPIRICAL EVALUATIONS

In this section, we evaluate the performance of our estimators using synthetic data generated from the well-known *autoregressive (AR)* time-series model of order 1 (see, e.g., [20]). Additionally, we examine the case where the autoregressive process deviates from the traditional AR model, featuring a non-Gaussian noise process. This latter setting serves to highlight the model-agnostic nature of our estimators.

A first-order autoregressive process, AR(1) can be defined as follows:

$$X_t = \phi_1 X_{t-1} + \epsilon_t, \quad t \geq 1$$

with parameter $\phi_1 \in \mathbb{R} \setminus \{0\}$ and an i.i.d. sequence of zero-mean normally distributed noise random variables $\epsilon_1, \epsilon_2, \ldots$ with variance σ^2. The process is stationary if $|\phi_1| < 1$ and, and by letting X_0 have distribution $\mathcal{N}(0, \sigma^2/(1 - \phi_1^2))$ we can guarantee stationarity of the process $X_0, X_1, X_2, \ldots$. To be able to calculate the estimation error, we compute the true β-mixing coefficients numerically. Furthermore, in the context of AR models, it is common to estimate the *autocorrelation function (ACF)* which is a measure of linear (as opposed to full) dependence. More specifically, the ACF at lag m is given by $\rho_m = E(X_0 X_m)$. A standard estimator of ρ_m when given a sample of length n can be obtained as

$$\widehat{\rho}_m = \frac{1}{(n-m)\widehat{\sigma}^2} \sum_{t=0}^{n-m} X_t X_{t+m} \tag{11}$$

with

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{t=0}^{n-1} X_t^2$$

In the case of an AR(1) process with Gaussian noise, an estimate of the β-mixing coefficients of the process can be readily obtained by using the ACF estimator given by (11). With this observation, we obtain a (model-specific) estimate of the β-mixing coefficient of the AR(1) process. This estimator is used as a point of comparison against our own estimators of β in this experiment. It is important to note that unlike the $\widehat{\rho}_m$-based estimator, our estimators do not have prior knowledge of the underlying time-series model, beyond assumptions on the smoothness of the densities. In fact, as part of our experiments, we further emphasize the model-agnostic nature of our estimator by comparing it with the ACF-based estimator under model-mismatch conditions: when the noise process is no longer Gaussian. In this setting, while the ACF-based estimator demonstrates bias and fails to consistently estimate the mixing coefficients, our estimator maintains remains consistent.

As discussed in Section IV, our estimators rely on KDEs to estimate β and as such require smoothness conditions on the the densities. Additionally, proper selection of $k^\star$, necessitates (loose) bounds on the mixing rate parameters of the process. Below, we discuss these conditions for AR(1) and provide insights into our approach for choosing $k^\star$ in our experiments.

KDE estimators for AR(1). Classical time-series models often assume Gaussian increments. This implies in various contexts that the joint densities of X_i and X_{i+m} is Gaussian. Gaussian densities are infinitely often differentiable and correspond to the smoothness parameter s of a Besov space approaching infinity. This yields a very fast rate of convergence of the KDE when the bandwidth h_n of the KDE is adapted to this context. In particular, the bandwidth h_n depends logarithmically on n and the rate of convergence will approach $\sqrt{\log(n)/n}$ in the L^2-distance and $\log(n)/n$ for the mean squared error independently of the dimension. A natural alternative to selecting the bandwidth in this way is to make use of popular heuristics for choosing h_n. In practice, it is common to use either *Scott's rule* or *Silverman's rule* to select h_n; in our experiments h_n is selected according to Scott's rule, which for an $\mathbb{R}^d$-valued sample sets $h_n = n^{-1/(4+d)}$. In the context of Besov spaces this corresponds to the case $s = 2$ [14, p.404] with corresponding rate of convergence of the KDE of order $n^{-2/(4+d)}$. Using Scott's rule leads to suboptimal rates of convergence of the KDE if our model assumptions are, in fact, correct since Scott's rule is pessimistic about the smoothness of the density and does not exploit the smoothness of the Gaussian distribution.

Choice of $k^\star$. Another hyperparameter we are required to select is $k^\star$, which corresponds the gap between time-points at which we collect observations. Larger values of $k^\star$ lead to observations that are closer to independent. The downside of large values of $k^\star$ is that significant amount of information is discarded. In Condition 2, we provided a general way to select $k^\star$ based on loose bounds on the mixing rate parameter. In the context of classical time-series models we have natural alternatives to this choice. In particular, for AR(1) models it is possible to bound the mixing-rate in terms of the AR(1) parameter and it is natural to replace our general assumption in Condition 2 by an AR-specific assumption on the parameters (e.g. the AR(1) parameter is smaller than 0.95). We develop these alternative approaches to select $k^\star$ as part of the model specific sections that follow.

As follows from [21, Thm.10.14], for an AR(1) process we can bound $\beta(m)$ from above and below using ρ_m as

$$\frac{1}{\pi}|\rho_m| - \frac{\rho_m^2 + (1/4)\rho_m^4}{(1 - \rho_m)^2}$$
$$\leq \beta(m)$$
$$\leq \frac{1}{\sqrt{2\pi}}|\rho_m| + \frac{\rho_m^2 + (1/16)\rho_m^4}{(1 - \rho_m)^2} \tag{12}$$

Note that the bound becomes tighter for small ρ_m and looser as ρ_m approaches 1. For larger values of ρ_m it indeed becomes trivial. The bound given in (12) allows us to control $\beta(m)$ by means of the correlation coefficients of the AR model. The correlation coefficients are in turn directly related to ϕ_1. It is natural to work in the AR(1) context with loose assumptions on ϕ_1 and the above discussion provides a path for turning such assumptions into bounds on $\beta(m)$ which we can then use to tune $k^\star$. The relationship between ρ_m, $m \geq 1$, and ϕ_1 is given by

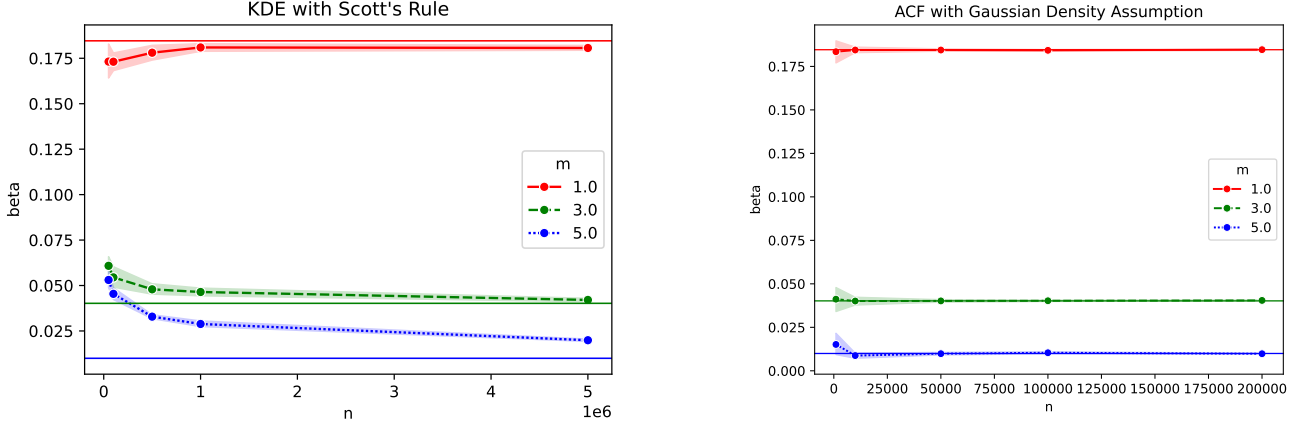$$\rho_m = \phi_1^m E(X_0^2)/E(X_0^2) = \phi_1^m.$$

Fig. 1. The plot on the left illustrates the performance of our KDE-based estimators, while the plot on the right demonstrates the performance of the ACF-based estimators for the estimation of $\beta(m)$, $m = 1, 3, 5$ when samples are generated according to AR(1). The horizontal lines represent the true values of β, with the top solid line (red) corresponding to $m = 1$, the middle solid line (green) to $m = 3$, and the bottom solid line (blue) to $m = 5$. The mean estimates (averaged over 20 rounds) are plotted for each m value, and the shaded areas indicate the uncertainty in these estimates. As n increases, the estimates clearly converge to their respective true β values. In the right plot, the results for the ACF-based estimator are shown. The ACF-based estimator takes advantage of the knowledge that the process is AR(1) and, more specifically, assumes Gaussian densities. This assumption leads to faster convergence in this particular setting, as the model assumption is accurate.

In particular, if we assume an upper bound b on $|\phi_1|$ then $|\rho_m| \leq b^m$. Substituting this into Equation (12) and ignoring second order terms provides us with the bound

$$\beta(m) \lesssim \frac{1}{\sqrt{2\pi}} b^m. \qquad (13)$$

Taking into account (13) when optimizing the bound on the expected error given by (33), and noting the smoothness of the Gaussian density, we propose choosing $k^\star$ in this case as

$$\left\lfloor \frac{\log \log(1/b) + (3/2) \log n}{\log(1/b)} \right\rfloor. \qquad (14)$$

Observe that for a fixed n, when $b \to 1$ we obtain $k^\star \to \infty$. Another alternative is to relax the requirement for the estimator to be based on (near)-independent samples by setting $k^\star = 0$. In this case, one can leverage the ergodicity of the underlying process to achieve consistency, as demonstrated in [11]. An advantage of this approach is that it eliminates the need for bounds on the mixing rate to select $k^\star$, allowing us to utilize the entire dataset without discarding any data. However, this might result in a higher variance, which would need to be appropriately controlled. Although theoretical results for finite-time analysis in the context of Markov processes and KDE-based estimators are currently lacking, we conjecture that it may be possible to prove the consistency of our estimator for $k^\star = 0$. We provide some simulation results below which lend support to this hypothesis.

Simulation Results. We start by considering an AR(1) model with $\phi_1 = 1/2$ and $\sigma^2 = 1$. Figure 1 shows the estimation results for the values of $\beta(m)$, $m = 1, 3$ and 5. The plot on the left shows the results for our estimator when $k^\star$ is chosen according to Equation (14) and using the assumption that $|\phi_1| \leq 0.9$. The plot on the right shows the results for a naive plug-in estimator that relies on the empirical ACF to calculate the empirical joint and marginal densities of the Gaussian

distribution, and integrate the absolute difference to estimate β. In this experiment, the true β values (numerically determined) are $\beta(1) \approx 0.1846, \beta(3) \approx 0.0402$ and $\beta(5) \approx 0.00996$. As expected, the ACF-based estimator which relies on the fact that the process is AR(1), is able to efficiently estimate β in this setting by calculating a simple 2×2 covariance matrix.

In a second set of experiments, we study the performance of the estimators under a mild model miss-specification. In this setting, the samples are given by $\exp(X_t) - E(\exp(X_t))$ where X_t is an AR(1) process. This yields a sequence of (centered) log-normal random variables. Observe that this transformation of X_t does not change the β-dependence of the original process. The results are shown in Figure 2. The plots on the top left and right show the results of our estimator where $k^\star$ is chosen according to (14) and $k^\star = 0$ respectively; the latter is more sample-efficient. Two notable observations can be gleaned from the results. 1. The model-mismatch encountered by the ACF-based estimator in this case results in a noticeable systematic bias in the estimates, even under a minor model mis-specification. This bias is particularly prominent for $m = 1$, where the dependencies are more significant. 2. Interestingly, in the case where $k^\star = 0$, the estimator performs well, despite the samples being dependent. This leads us to believe that it may be possible to prove consistency of our KDE-estimators without the need to rely on blocks that are $k^\star$ steps apart.

VI. DISCUSSION

We have introduced novel methods for estimating the β-mixing coefficients of a real-valued geometrically ergodic Markov process from a sample-path of length n. Under standard smoothness conditions on the densities, we have established a convergence rate of $\mathcal{O}(\log(n) n^{-[s]/(2[s]+2)})$ for the expected
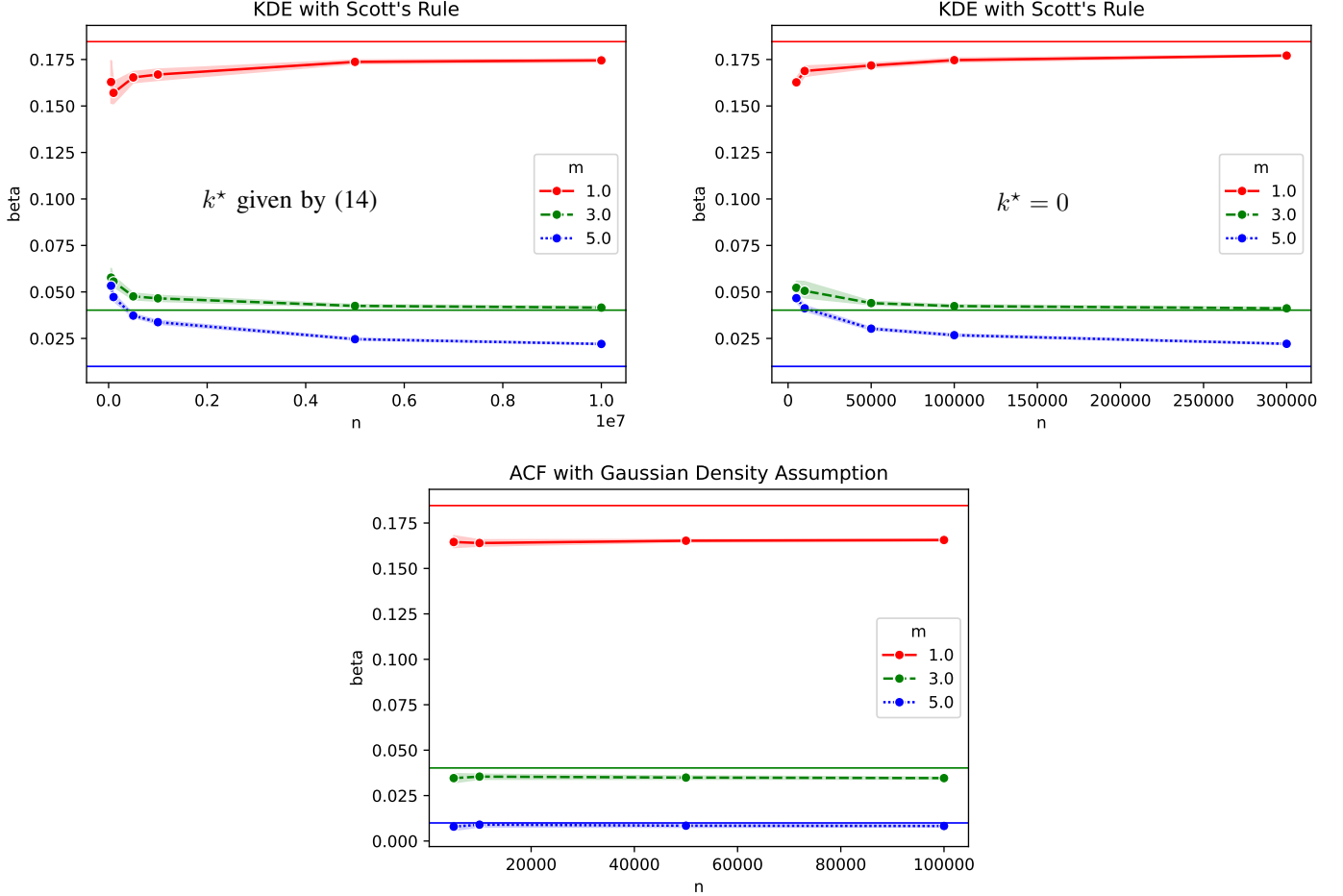
Fig. 2. Performance of our estimators (top) and the ACF-based estimator (bottom) when the sample is a transformation of an AR(1) process.

error of our estimator, and have provided a high-probability bound on the estimation error. Furthermore, we have derived analogous bounds in the case of a finite state-space, where no density assumptions are required, and demonstrate an expected error rate of $\mathcal{O}(\log(n)n^{-1/2})$. Although our work is primarily theoretical, we have also presented empirical evaluations to further validate the performance of the proposed estimators. In our experiments, we demonstrate the efficacy of our estimators across a range of geometircally ergodic Markov processes without being tied to any particular modeling specification, such as linear or autoregressive structures.

Our estimators rely on (near)-independent samples obtained from blocks separated by some $k^\star$ time-steps in the sample path. While this is a classical technique, based on Berbee's coupling lemma, allowing to control dependencies, it has its downsides. First, to determine an appropriate value for $k^\star$ we require bounds, albeit loose, on the mixing parameters. Moreover, using blocks that are $k^\star$ steps apart leads to sample-inefficiency. An alternative strategy could involve eliminating the need for (near)-independent samples by setting $k^\star = 0$. As in [11] it may be possible to exploit the underlying process's ergodicity to ensure consistency. However, this might result in a higher variance to be controlled. While to the best of our knowledge there are currently no theoretical results available

for finite-time analysis in the context of Markov processes and KDE-based estimators, we conjecture that it could be feasible to demonstrate the consistency of our estimator under the condition that $k^\star = 0$. We leave this investigation for future work.

VII. TECHNICAL PROOFS

In this section, we provide proofs for our theorems. A common ingredient is a coupling argument for time-series, which allows one to move from dependent samples to independent blocks. This is facilitated by Lemma 7 below, which is a standard result based on, commonly used in the analysis of dependent time-series, see e.g. [22]–[24]. For completeness, we provide a proof of this lemma, which in turns relies on a coupling Lemma of [25] stated below.

Lemma 7. *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 8 ([25]). *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 7. 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 8 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 8, 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

$$\sigma(Y_0^*, Y_1^*, \ldots, Y_{j-1}^*) \subseteq \mathcal{U}_j \vee \mathcal{V}_j \tag{15}$$

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,

$$\begin{aligned}
P(U \cap W | \mathcal{V}_j) &= \mathbb{E}(\mathbf{1}_U \mathbf{1}_W | \mathcal{V}_j) \\
&= \mathbb{E}(\mathbb{E}(\mathbf{1}_U \mathbf{1}_W | \mathcal{V}_j \vee \sigma(Y_j)) | \mathcal{V}_j) \tag{16} \\
&= \mathbb{E}(\mathbf{1}_W \mathbb{E}(\mathbf{1}_U | \mathcal{V}_j \vee \sigma(Y_j)) | \mathcal{V}_j) \tag{17} \\
&= \mathbb{E}(\mathbf{1}_W \mathbb{E}(\mathbf{1}_U) | \mathcal{V}_j) \tag{18} \\
&= P(U) P(W | \mathcal{V}_j) \\
&= P(U | \mathcal{V}_j) P(W | \mathcal{V}_j) \tag{19}
\end{aligned}$$

where (16) follows from the fact that $\mathcal{V}_j \subseteq \sigma(Y_j) \vee \mathcal{V}_j$, (17) follows from noting that $W \in \sigma(Y_j)$, (18) holds since $\mathcal{U}_j$ is independent of $\sigma(Y_0, \ldots, Y_j)$, and (19) follows since $\mathcal{U}_j$ is independent of $\mathcal{V}_j$. Therefore, and $\mathcal{U}_j$, $\mathcal{V}_j$ and $\sigma(Y_j)$ form a Markov triplet in the sense of [7, Vol. 1 Definition 7.1 pp. 205]. Thus, as follows from [7, Vol. 1 Theorem 7.2 pp. 205] we obtain,

$$\beta(\mathcal{U}_j \vee \mathcal{V}_j, \sigma(Y_j)) = \beta(\mathcal{V}_j, \sigma(Y_j)). \tag{20}$$

In light of (15) and (20), and noting that by construction $\beta(\mathcal{V}_j, \sigma(Y_j)) \leq \beta(k)$ we obtain

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

□

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 (34).* Enlarge Ω if necessary in order for Lemma 7 to be applicable. As follows from Lemma 7 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

$$\Pr(\{\exists i \in 0, \ldots, N-1 : Z_i^* \neq Z_i\}) \leq N\beta(k) \tag{21}$$

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 (6) and let

$$\widehat{\beta}_N^*(m) = \frac{1}{2} \int_{\mathbb{R}^2} |\widehat{f}_{m,N}^* - \widehat{f}_{0,N}^* \otimes \widehat{f}_{0,N}^*| d\lambda_2 \tag{22}$$

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

$$\begin{aligned}
&\|f_0 \otimes f_0 - \widehat{f}_{0,N}^* \otimes \widehat{f}_{0,N}^*\|_1 \\
&= \int_x \int_y |f_0(x)f_0(y) - \widehat{f}_{0,N}^*(x)\widehat{f}_{0,N}^*(y)| d\lambda(x) d\lambda(y) \\
&= \int_x \int_y \Big| f_0(x)f_0(y) - \widehat{f}_{0,N}^*(x)f_0(y) \\
&\qquad + \widehat{f}_{0,N}^*(x)f_0(y) - \widehat{f}_{0,N}^*(x)\widehat{f}_{0,N}^*(y)\Big| d\lambda(x) d\lambda(y) \\
&\leq \int_y f_0(y) \int_x |f_0(x) - \widehat{f}_{0,N}^*(x)| d\lambda(x) d\lambda(y) \\
&\qquad + \int_x |\widehat{f}_{0,N}^*(x)| \int_y |f_0(y) - \widehat{f}_{0,N}^*(y)| d\lambda(y) d\lambda(x) \\
&\leq (1 + c_0) \|f_0 - \widehat{f}_{0,N}^*\|_1 \tag{23}
\end{aligned}$$

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

$$\begin{aligned}
&|\beta(m) - \widehat{\beta}_N^*(m)| \\
&= \frac{1}{2} \left| \int |f_m - f_0 \otimes f_0| d\lambda_2 - \int |\widehat{f}_{m,N}^* - \widehat{f}_{0,N}^* \otimes \widehat{f}_{0,N}^*| d\lambda_2 \right| \\
&\leq \frac{1}{2} \int |f_m - \widehat{f}_{m,N}^*| d\lambda_2 + \frac{1}{2} \int |f_0 \otimes f_0 - \widehat{f}_{0,N}^* \otimes \widehat{f}_{0,N}^*| d\lambda_2 \\
&= \frac{1}{2} \|f_m - \widehat{f}_{m,N}^*\|_1 + \frac{1}{2} \|f_0 \otimes f_0 - \widehat{f}_{0,N}^* \otimes \widehat{f}_{0,N}^*\|_1 \\
&\leq \frac{1}{2} \left(\|f_m - \widehat{f}_{m,N}^*\|_1 + (1 + c_0)\|f_0 - \widehat{f}_{0,N}^*\|_1 \right) \tag{24}
\end{aligned}$$

where (24) follows from (23). Next, it is straightforward to check that if $f_m \in B_{1\infty}^s(\mathbb{R}^2)$ with $\|f\|_{B_{1\infty}^s(\mathbb{R}^2)} \leq \Lambda$, then $f_0 = \int_\mathbb{R} f_m d\lambda \in B_{1\infty}^s(\mathbb{R})$ with $\|f\|_{B_{1\infty}^s(\mathbb{R})} \leq \Lambda$. Moreover,

observe that, as follows from Remark 1, for all $f \in B_{1\infty}^s(\mathbb{R})$ and $g \in B_{1\infty}^s(\mathbb{R}^2)$ we have $\|f\|_{W_1^{[s]}(\mathbb{R})} \leq \|f\|_{B_{1,\infty}^s(\mathbb{R})}$ and $\|g\|_{W_1^{[s]}(\mathbb{R}^2)} \leq \|g\|_{B_{1,\infty}^s(\mathbb{R}^2)}$. Therefore, with the choice of bandwidth h specified in the theorem statement, by [14, Proposition 4.1.5 and Proposition 4.3.33] and an argument analogous to that of [14, Proposition 5.1.7], we obtain,

$$\sup_{f_0: \|f\|_{B_{1\infty}^s(\mathbb{R})} \leq \Lambda} \mathbb{E}\|\widehat{f}_{0,N}^* - f_0\|_1 \leq \widetilde{C} N^{-\frac{[s]}{2[s]+2}} \quad (25)$$

where $\widetilde{C} = 2(L_1)^{\frac{[s]}{[s]+1}}(c\Lambda)^{\frac{1}{[s]+1}}$. Similarly, by [14, pp. 404] we have,

$$\sup_{f_m: \|f\|_{B_{1\infty}^s(\mathbb{R}^2)} \leq \Lambda} \mathbb{E}\|\widehat{f}_{m,N}^* - f_m\|_1 \leq \widetilde{C} N^{-\frac{[s]}{2[s]+2}} \quad (26)$$

Set $C := (2 + c_0)\widetilde{C}/2$. Define the event

$$E := \{Z_i^* = Z_i, \ i \in 0, \ldots, N-1\}.$$

We obtain,

$$\mathbb{E}|\beta(m) - \widehat{\beta}_N(m)|$$
$$\leq \mathbb{E}|\beta(m) - \widehat{\beta}_N^*(m)|$$
$$\quad + \mathbb{E}[|\widehat{\beta}_N^*(m) - \widehat{\beta}_N(m)| \, |E^c] \Pr(E^c) \quad (27)$$
$$\leq \mathbb{E}|\beta(m) - \widehat{\beta}_N^*(m)| + 2N\beta(k) \quad (28)$$
$$\leq CN^{-\frac{[s]}{2[s]+2}} + 2N\beta(k) \quad (29)$$
$$\leq CN^{-\frac{[s]}{2[s]+2}} + 2N\eta e^{-\gamma k} \quad (30)$$
$$\leq C\left(\frac{4k}{n-4k}\right)^{\frac{[s]}{2[s]+2}} + \frac{n}{k}\eta e^{-\gamma k} \quad (31)$$
$$\leq C\left(\frac{8k}{n}\right)^{\frac{[s]}{2[s]+2}} + \frac{n}{k}\eta e^{-\gamma k} \quad (32)$$
$$\leq \frac{8Ck}{n^{\frac{[s]}{2[s]+2}}} + \eta n e^{-\gamma k} \quad (33)$$

where (27) follows from triangle inequality and observing that under E the estimators $\widehat{\beta}_N$ and $\widehat{\beta}_N^*$ are equal, (28) follows from (21), (29) follows from (24),(25) and (26), (30) follows from observing that $\beta(k) \leq 1$ and the geometric ergodicity of the process, and (31) and (32) follow from the definition of N and the fact that $2 \leq k \leq \lfloor n/8 \rfloor$. Optimizing (33) for k we obtain

$$k^\star = \left\lfloor \frac{1}{\gamma}\left(\log \frac{\gamma\eta}{8C} + \left(\frac{3[s]+2}{2[s]+2}\right)\log n\right)\right\rfloor \quad (34)$$

which in turn leads to

$$\mathbb{E}|\beta(m) - \widehat{\beta}_N^*(m)|$$
$$\leq \frac{8Cn^{-\frac{[s]}{2[s]+2}}}{\gamma}\left(e^\gamma + \log \frac{\gamma\eta}{8C} + \left(\frac{3[s]+2}{2[s]+2}\right)\log n\right) \quad (35)$$

This completes the proof of the bound on the expected error. For $t > 0$ define the event $U_t := \{N\|\widehat{f}_{m,N}^* - \mathbb{E}(\widehat{f}_{m,N}^*)\|_1 \geq (3/2)N\mathbb{E}\|\widehat{f}_{m,N}^* - \mathbb{E}(\widehat{f}_{m,N}^*)\|_1 + \sqrt{2Nt}\|K\|_1 + t5\|K\|_1\}$. To obtain the high probability bound observe that [14, Theorem 5.1.13] states that

$$\Pr(U) \leq e^{-t} \quad (36)$$

for all $t > 0$. 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,

$$\|\widehat{f}_{m,N}^* - f_m\|_1$$
$$\leq \|\widehat{f}_{m,N}^* - \mathbb{E}(\widehat{f}_{m,N}^*)\|_1 + \|K_{h_N} * f_m - f_m\|_1. \quad (37)$$

The latter term can be bounded by using [14, Proposition 4.3.33],

$$\|K_{h_N} * f_m - f_m\|_1 \leq h_N^{[s]}\|f_m\|_{W_1^{[s]}(\mathbb{R}^2)}, \quad (38)$$

Note that there is a typo in [14, Proposition 4.3.33] which states the result as in the one-dimensional case. The inequality (38) relies on the remainder term of a Taylor series. The remainder term in $\|\cdot\|_1$-norm is upper bounded by (Minkowski inequality for integrals)

$$[s]h_N^{[s]}\sum_{|\alpha|=[s]}\frac{1}{\alpha!}\int_0^1 (1-u)^{[s]-1}\|D^\alpha f_m\|_1 \, du$$
$$\leq [s]h_N^{[s]}\int_0^1 (1-u)^{[s]-1}du\|f_m\|_{W_1^{[s]}(\mathbb{R}^2)}, \quad (39)$$

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 $1 - e^{-u}$,

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

Recall that

$$\|f_m\|_{W_1^{[s]}(\mathbb{R}^2)} \leq \|f_m\|_{B_{1,\infty}^2(\mathbb{R}^2)}$$

and, therefore, from (26) it follows that for any f_m such that $\|f_m\|_{B_{1,\infty}^2(\mathbb{R}^2)} \leq \Lambda$, with probability $1 - e^{-u}$ we have,

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

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

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

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 (23) that

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

Substituting this into (24) gives that with probability at least $1 - 2e^{-u}$,

$$|\beta(m) - \widehat{\beta}_N^*(m)|/(1 + c_0/2)$$
$$\leq \sqrt{2u/N}\|K\|_1 + u5\|K\|_1/N \qquad (40)$$
$$+ \left((3/2)\widetilde{C} + \Lambda(c\Lambda)^{-\frac{[s]^2}{[s]+1}}\left(\frac{[s]-1}{2}\right)^{-\frac{[s]^2}{2[s]+2}}\right)N^{-\frac{[s]}{2[s]+2}}$$

Furthermore, with probability

$$1 - N\beta(k^\star) \geq 1 - 8Cn^{-\frac{[s]}{2[s]+2}}$$

it holds that $\widehat{f}_{0,N}^* = \widehat{f}_{0,N}$ and $\widehat{f}_{m,N}^* = \widehat{f}_{m,N}$. By setting

$$u = \log(n)[s]/(2[s]+2)$$

in (40) with probability at least $1 - 8Cn^{-\frac{[s]}{2[s]+2}}$ we have

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

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

$$|\beta(m) - \widehat{\beta}_N(m)|$$
$$\leq 16(1 + \frac{c_0}{2})\left(\frac{3}{2}\widetilde{C} + \Lambda(c\Lambda)^{-\frac{[s]^2}{[s]+1}}\left(\frac{[s]-1}{2}\right)^{-\frac{[s]^2}{2[s]+2}}\right.$$
$$+ \left.\left(\sqrt{\frac{[s]}{([s]+1)}} + \frac{5[s]}{(2[s]+2)}\right)\log(n)\|K\|_1\right)\left(\frac{n}{4k^\star}\right)^{-\frac{[s]}{2[s]+2}}$$

Noting that $(4k^\star)^{\frac{[s]}{2[s]+2}} \leq 4k^\star$, defining C_1 as

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

and letting

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

we obtain

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

$\square$

Proof of Theorem 5. 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 (45). As follows from Lemma 7 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 $\mathcal{X}^2$ and has the same distribution as Z_i such that

$$\Pr\left(\{\exists i \in 0, \ldots, N - 1 : Z_i^* \neq Z_i\}\right) \leq N\beta(k) \qquad (41)$$

Define $\widehat{\beta}_N^*(m)$ as

$$\frac{1}{2}\sum_{u\in\mathcal{X}}\sum_{v\in\mathcal{X}}|\widehat{P}_{m,N}^*(\{(u,v)\}) - \widehat{P}_{0,N}^*(\{u\})\widehat{P}_{0,N}^*(\{v\})|$$

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 $z \in \mathcal{X} \times \mathcal{X}$ we have,

$$\mathbb{E}[|\widehat{P}_{m,N}^*(z) - P_m(z)|]$$
$$\leq \frac{1}{N}\left(\sum_{i=0}^{N-1}\mathbb{E}(\mathbf{1}_{\{z\}}(Z_i^*) - \mathbb{E}\mathbf{1}_{\{z\}}(Z_i^*))^2\right)^{1/2}$$
$$\leq \sqrt{P_m(z)/N} \qquad (42)$$

where the second inequality is due to

$$\mathbb{Var}\left(\mathbf{1}_{\{z\}}(Z_i^*) - \mathbb{E}\mathbf{1}_{\{z\}}(Z_i^*)\right) \leq P_m(z)$$

Similarly, for each $u \in \mathcal{X}$ we obtain

$$\mathbb{E}[|\widehat{P}_{0,N}^*(u) - P_0(u)|] \leq \sqrt{P_0(u)/N}$$

It follows that

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

In much the same way as in the proof of Theorem 3, let

$$E := \{Z_i^* = Z_i, \ i \in 0, \ldots, N - 1\}$$

and observe that

$$\mathbb{E}(|\widehat{\beta}_N^*(m) - \widehat{\beta}_N(m)| \,|E) = 0$$

Moreover, recall that $k \leq \lfloor n/8 \rfloor$. We obtain,

$$\mathbb{E}|\beta(m) - \widehat{\beta}_N(m)| \leq \mathbb{E}|\beta(m) - \widehat{\beta}_N^*(m)| + 2N\beta(k) \quad (43)$$

$$\leq |\mathcal{X}|\left(\frac{4k}{n-4k}\right)^{1/2} + 2n\eta e^{-\gamma k}$$

$$\leq |\mathcal{X}|\left(\frac{8k}{n}\right)^{1/2} + 2n\eta e^{-\gamma k}$$

$$\leq \sqrt{8}|\mathcal{X}|n^{-1/2}k + 2n\eta e^{-\gamma k} \quad (44)$$

where (28) follows from (41). Optimizing (44) we obtain

$$k^\star = \left\lfloor \frac{1}{\gamma}\log\left(\frac{\eta\gamma n^{3/2}}{\sqrt{2}|\mathcal{X}|}\right)\right\rfloor \quad (45)$$

where n is taken large enough so ensure that $k^\star \geq 2$ (see (46) below). This choice of $k^\star$ and n leads to

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

Next, take $N = N(k^\star, n) = \lfloor \frac{n-k^\star}{2(k^\star+1)}\rfloor$, with $k^\star$ given by (45) and let n be large enough so that

$$n^{3/2} \geq \max\left\{\frac{\sqrt{2}|\mathcal{X}|e^{2\gamma}}{\eta\gamma}, \frac{\eta\gamma}{|\mathcal{X}|}\right\}. \quad (46)$$

Substituting for $k^\star$ and noting that

$$N = \left\lfloor\frac{n-k^\star}{2(k^\star+1)}\right\rfloor \leq \frac{\gamma n}{2\log\left(\frac{\eta\gamma n^{3/2}}{\sqrt{2}|\mathcal{X}|}\right)}$$

we have,

$$N\eta e^{-\gamma k^\star} \leq \frac{e\sqrt{2}|\mathcal{X}|n^{-1/2}}{2\log\left(\frac{\eta\gamma n^{3/2}}{\sqrt{2}|\mathcal{X}|}\right)}. \quad (47)$$

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

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

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

$$\Pr(|\widehat{P}_{m,N}^*(\{(u,v)\}) - P_m(\{(u,v)\})| \geq \frac{\epsilon}{2|\mathcal{X}|^2})$$
$$\leq 2\exp\left\{-\frac{N\epsilon^2}{2|\mathcal{X}|^4}\right\}$$

It follows that

$$\Pr(|\widehat{\beta}_N^*(m) - \beta(m)| \geq \epsilon/2)$$
$$\leq \sum_{u,v}\Pr(|P_m(\{(u,v)\}) - \widehat{P}_{m,N}^*(\{(u,v)\})| \geq \frac{\epsilon}{2|\mathcal{X}|^2})$$
$$\quad + 2\sum_u\Pr(|P_0(\{v\}) - \widehat{P}_{0,N}^*(\{u\})| \geq \frac{\epsilon}{2|\mathcal{X}|})$$
$$\leq 2|\mathcal{X}|^2\exp\left\{-\frac{N\epsilon^2}{2|\mathcal{X}|^4}\right\} + 2|\mathcal{X}|\exp\left\{-\frac{N\epsilon^2}{2|\mathcal{X}|^2}\right\}$$
$$\leq 4|\mathcal{X}|^2\exp\left\{-\frac{N\epsilon^2}{2|\mathcal{X}|^4}\right\}$$
$$\leq 4|\mathcal{X}|^2\exp\{-\frac{(n-4k^\star)\epsilon^2}{8k^\star|\mathcal{X}|^4}\} \quad (48)$$
$$\leq 4|\mathcal{X}|^2\exp\{-\frac{n\epsilon^2}{16k^\star|\mathcal{X}|^4}\} \quad (49)$$
$$= 4|\mathcal{X}|^2\exp\left\{-\frac{\gamma n\epsilon^2}{16|\mathcal{X}|^4\left(\log\left(\frac{\eta\gamma}{\sqrt{2}|\mathcal{X}|}\right) + \frac{3}{2}\log n\right)}\right\} \quad (50)$$
$$\leq 4|\mathcal{X}|^2\exp\{-\frac{\gamma n\epsilon^2}{48|\mathcal{X}|^4\log n}\} \quad (51)$$

where, (48) follows from the choice of $N = \lfloor\frac{n-k^\star}{2(k^\star+1)}\rfloor$ and noting that $k^\star \geq 2$, (49) follows from recalling that in general, k (and thus also $k^\star$), is less than $\lfloor n/8 \rfloor$, and finally, (50) and (51) follow from substituting the value of $k^\star$ as given by (34) and observing that by (46) we have $\frac{3}{2}\log n \geq \log(\frac{\eta\gamma}{|\mathcal{X}|})$. Hence, by (41), (47) and (51) we obtain,

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

$\square$

Proof of Theorem 6. As in the proof of Theorem 5, 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 $X_0, \ldots, X_n$ define

$$\widetilde{Z}_i = (X_{2i(k+1)}, X_{2i(k+1)+1}, \ldots, X_{(2i+1)k+2i})$$

for $i = 0, 1, \ldots, N-1$ where $N = N(k,n) := \lfloor\frac{n-k}{2(k+1)}\rfloor$ for some fixed $k \in 1, \ldots, \lfloor n/8 \rfloor$; an optimal choice for k is specified later in the proof, see (60). By Lemma 7 there exists a sequence of independent random vectors $\widetilde{Z}_i^* = (\widetilde{Z}_{i,0}^*, \ldots, \widetilde{Z}_{i,k}^*)$ for $i = 0, 1, \ldots, N-1$ each of which takes value in $\mathcal{X}^{k+1}$ and has the same distribution as $\widetilde{Z}_i$ such that

$$\Pr\left(\{\exists i \in 0, \ldots, N-1 : \widetilde{Z}_i^* \neq \widetilde{Z}_i\}\right) \leq N\beta(k). \quad (52)$$

Define

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

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 5, for each $u \in \mathcal{X}$ it holds that

$$\mathbb{E}[|\widehat{P}^\dagger_{0,N}(u) - P_0(u)|] \leq \sqrt{P_0(u)/N} \leq N^{-1/2}. \tag{53}$$

Define the class of indicator functions

$$\mathcal{H}_k = \{h_{m,z} : z \in \mathcal{X} \times \mathcal{X}, \ m = 1, \ldots, k\}.$$

where

$$h_{m,z} : \mathcal{X}^{k+1} \to \{0,1\}$$
$$h_{m,z}(\mathbf{x}) := \mathbf{1}_{\{z\}}(x_0, x_m).$$

One can verify that the VC-dimension of $\mathcal{H}_k$ is at most $\log_2(|\mathcal{X}|k)$. To see this, first that the indicator functions in $\mathcal{H}_k$ depend on only two coordinates at a time, rather than on all coordinates simultaneously. Consequently, the VC-dimension of $\mathcal{H}_k$ will typically exceed one but remain limited by the combinatorial structure of these pairwise tests. The upper bound follows from the standard result that the VC-dimension of a finite hypothesis class $\mathcal{H}_k$ is at most $\log_2 |\mathcal{H}_k|$. In our case, the class $\mathcal{H}_k$ has cardinality $|\mathcal{X}|^2 k$. However, since the first coordinate can be fixed to a constant value (say, 1), we can restrict attention to a subclass of functions of size $|\mathcal{X}|k$ that can shatter as many points as $\mathcal{H}_k$. Therefore, the VC-dimension of $\mathcal{H}_k$ is bounded by $\log_2(|\mathcal{X}|k)$.

Therefore, as follows from [14, pp. 217] it holds that,

$$\mathbb{E}[\sup_{\substack{m \in 1,\ldots,k \\ (u,v) \in \mathcal{X}^2}} |P^\dagger_{m,N}((u,v)) - P_m((u,v))|]$$

$$\leq \sqrt{\frac{8 \log_2(|\mathcal{X}|k) \log N}{N}} \tag{54}$$

By (53) and (54) we have,

$$2\mathbb{E}[\sup_{m \in 1,\ldots,k} |\widehat{\beta}^\dagger_N(m) - \beta(m)|]$$

$$\leq \mathbb{E}[\sup_{m \in 1,\ldots k} \sum_{(u,v) \in \mathcal{X}^2} |P_m(\{(u,v)\}) - \widehat{P}^\dagger_{m,N}(\{(u,v)\})|$$

$$+ 2\sum_{u \in \mathcal{X}} |P_0(\{u\}) - \widehat{P}^\dagger_{0,N}(\{u\})|]$$

$$\leq |\mathcal{X}|^2 \mathbb{E}[\sup_{\substack{m \in 1,\ldots,k \\ (u,v) \in \mathcal{X}^2}} |P_m(\{(u,v)\}) - \widehat{P}^\dagger_{m,N}(\{(u,v)\})|]$$

$$+ 2\sum_{u \in \mathcal{X}} \mathbb{E}|P_0(\{u\}) - \widehat{P}^\dagger_{0,N}(\{u\})|$$

$$\leq |\mathcal{X}|^2 \sqrt{\frac{8 \log_2(|\mathcal{X}|k) \log N}{N}} + 2|\mathcal{X}|N^{-1/2}$$

$$\leq 2|\mathcal{X}|^2 \sqrt{\frac{8 \log_2(|\mathcal{X}|k) \log N}{N}} \tag{55}$$

Take $n \geq \frac{2k(k+1)}{3k+2}$. From the coupling argument given earlier we obtain,

$$\mathbb{E}[\sup_{m \in 1,\ldots,k} |\widehat{\beta}_N(m) - \beta(m)|]$$

$$\leq \mathbb{E}[\sup_{m \in 1,\ldots,k} |\widehat{\beta}^\dagger_N(m) - \beta(m)|] + N\beta(k)$$

$$\leq |\mathcal{X}|^2 \sqrt{\frac{8 \log_2(|\mathcal{X}|k) \log N}{N}} + N\eta e^{-\gamma k} \tag{56}$$

$$\leq |\mathcal{X}|^2 \sqrt{\frac{8}{N}} \log_2(N|\mathcal{X}|) + N\eta e^{-\gamma k} \tag{57}$$

$$\leq 4|\mathcal{X}|^2 \sqrt{\frac{2k}{n}} \log_2(n|\mathcal{X}|) + n\eta e^{-\gamma k} \tag{58}$$

$$\leq 4|\mathcal{X}|^2 \sqrt{\frac{2}{n}} \log_2(n|\mathcal{X}|)k + n\eta e^{-\gamma k} \tag{59}$$

where (56) follows from (55), and (57) follows from noting that $N = N(k,n) \geq k$ for $n \geq \frac{2k(k+1)}{3k+2}$; similarly, (58) follows form the choice of $N = N(k,n) \leq \frac{n-k}{2(k+1)} - 1$ and noting that $k \geq 1$. Optimizing (59) we have

$$k^\dagger = \left\lfloor \frac{1}{\gamma} \log \left(\frac{\eta \gamma n^{3/2}}{4\sqrt{2}|\mathcal{X}|^2 \log_2(n|\mathcal{X}|)} \right) \right\rfloor \tag{60}$$

with

$$n \geq \max \left\{ \frac{4\sqrt{2}|\mathcal{X}|^3 e^\gamma}{\eta \gamma}, \frac{2k(k+1)}{3k+2} \right\}. \tag{61}$$

This leads to

$$\mathbb{E}[\sup_{m \in 1,\ldots,k} |\widehat{\beta}_N(m) - \beta(m)|]$$

$$\leq \frac{4\sqrt{2}|\mathcal{X}|^2 n^{-1/2} \log_2(n|\mathcal{X}|)}{\gamma}$$

$$\times \left(e + \log \left(\frac{\eta \gamma n^{3/2}}{4\sqrt{2}|\mathcal{X}|^2 \log_2(n|\mathcal{X}|)} \right) \right).$$

Take $N = N(k^\dagger, n) = \lfloor \frac{n-k^\dagger}{2(k^\dagger+1)} \rfloor$, with $k^\dagger$ given by (60). It follows that,

$$N\eta e^{-\gamma k^\dagger} \leq \frac{2e\sqrt{2}|\mathcal{X}|^2 \log_2(n|\mathcal{X}|)n^{-1/2}}{\log(\frac{\eta \gamma n^{3/2}}{4\sqrt{2}|\mathcal{X}|^2 \log_2(n|\mathcal{X}|)})} \tag{62}$$

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

$$\Pr(|\widehat{P}^\dagger_{0,N}(\{u\}) - P_0(\{u\})| \geq \frac{\epsilon}{2|\mathcal{X}|}) \leq 2\exp\left\{-\frac{N\epsilon^2}{2|\mathcal{X}|^2}\right\} \tag{63}$$

Furthermore, noting that $\mathcal{H}_{k^\dagger}$ is a VC-class, by [26, Theorem 12.5] for $\epsilon > 0$ we have,

$$\Pr(\sup_{\substack{m \in 1,\ldots,k^\dagger \\ (u,v) \in \mathcal{X}^2}} |\widehat{P}^\dagger_{m,N}(\{(u,v)\}) - P_m(\{(u,v)\})| \geq \frac{\epsilon}{2|\mathcal{X}|^2})$$

$$\leq 8 \log_2(|\mathcal{X}|k)e^{-\frac{N\epsilon^2}{128|\mathcal{X}|^4}} \tag{64}$$

Furthermore, noting that in general all k are taken to be less than $\lfloor n/8 \rfloor$, a minor manipulation gives,

$$
\begin{aligned}
N &\geq \frac{n-k}{2(k+1)} - 1 \\
&\geq \frac{n-4k}{4k} \qquad\qquad \text{since } k \geq 2 \\
&\geq \frac{n}{8k} \qquad\qquad\qquad\qquad\qquad (65)
\end{aligned}
$$

where (65) is from noting that $k \leq \lfloor \frac{n}{8} \rfloor$. Therefore, for any $\epsilon > 0$ we have,

$$
\frac{1}{|\mathcal{X}|^2} \Pr(\sup_{m \in 1,\dots,k^\dagger} |\widehat{\beta}_N^\dagger(m) - \beta(m)| \geq \epsilon/2)
$$

$$
\leq \Pr\left(\sup_{\substack{m \in 1,\dots,k^\dagger \\ (u,v) \in \mathcal{X}^2}} |P_m(\{(u,v)\}) - \widehat{P}_{m,N}^\dagger(\{(u,v)\})| \geq \frac{\epsilon}{2|\mathcal{X}|^2} \right)
$$

$$
+ \frac{2}{|\mathcal{X}|^2} \sum_u \Pr\left(|P_0(\{v\}) - \widehat{P}_{0,N}^\dagger(\{u\})| \geq \frac{\epsilon}{2|\mathcal{X}|} \right)
$$

$$
\leq 8 \log(|\mathcal{X}| k^\dagger) \exp\left\{ -\frac{N\epsilon^2}{128|\mathcal{X}|^4} \right\} + \frac{2}{|\mathcal{X}|} \exp\left\{ -\frac{N\epsilon^2}{2|\mathcal{X}|^2} \right\}
$$

$$
\leq 16 \log(|\mathcal{X}| k^\dagger) \exp\left\{ -\frac{N\epsilon^2}{128|\mathcal{X}|^4} \right\}
$$

$$
\leq 16 \log(|\mathcal{X}| k^\dagger) \exp\left\{ -\frac{n\epsilon^2}{1024 k^\dagger |\mathcal{X}|^4} \right\} \qquad (66)
$$

$$
\leq 16 \log\left(\frac{3|\mathcal{X}|}{2\gamma} \log(\eta^{2/3}\gamma^{2/3} n) \right)
$$

$$
\times \exp\left\{ -\frac{\gamma n\epsilon^2}{3072|\mathcal{X}|^4 \log n} \right\} \qquad (67)
$$

where, the second inequality follows from (63) and (64), (66) follows from the choice (65), and (67) follows from substituting for the value of $k^\dagger$ as given by (60) and observing that by (61) we have

$$
\frac{n^{3/2}}{\log(n|\mathcal{X}|)} \geq \frac{n}{|\mathcal{X}|} \geq \frac{\eta\gamma}{8\sqrt{2}|\mathcal{X}|^2}.
$$

Hence, by (52), (62) and (67) we obtain,

$$
\Pr(\sup_{m=1,\dots,k^\dagger} |\widehat{\beta}_N(m) - \beta(m)| \geq \epsilon)
$$

$$
\leq N\eta e^{-\gamma k^\dagger} + \Pr(\sup_{m=1,\dots,k^\dagger} |\widehat{\beta}_N^\dagger(m) - \beta(m)| \geq \epsilon/2)
$$

$$
\leq \frac{4\sqrt{2}|\mathcal{X}|^2 \log(n|\mathcal{X}|) n^{-1/2}}{\log(\frac{\eta\gamma n^{3/2}}{8\sqrt{2}|\mathcal{X}|^2})}
$$

$$
+ 16|\mathcal{X}|^2 \log\left(\frac{3|\mathcal{X}|}{2\gamma} \log(\eta^{2/3}\gamma^{2/3} n) \right)
$$

$$
\times \exp\left\{ -\frac{\gamma n\epsilon^2}{3072|\mathcal{X}|^4 \log n} \right\}
$$

and the result follows. $\qquad\square$

References

[1] I. A. Ibragimov, "Some limit theorems for stationary processes," *Theory of Probability & its Applications*, vol. 7, no. 4, pp. 349–382, 1962.

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

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

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

[5] E. Rio, "Inégalités de hoeffding pour les fonctions lipschitziennes de suites dépendantes," *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, vol. 330, no. 10, pp. 905–908, 2000.

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

[7] R. C. Bradley, *Introduction to Strong Mixing Conditions*. Kendrick Press, 2007, vol. 1,2, 3.

[8] A. Kontorovich and K. Ramanan, "Concentration inequalities for dependent random variables via the martingale method," *The Annals of Probability*, vol. 36, no. 6, pp. 2126 – 2158, 2008.

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

[10] A. B. Nobel, "Hypothesis testing for families of ergodic processes," *Bernoulli*, vol. 12, no. 2, pp. 251–269, 2006.

[11] A. Khaleghi and G. Lugosi, "Inferring the mixing properties of a stationary ergodic process from a single sample-path," *IEEE Transactions on Information Theory*, vol. 69, no. 6, pp. 4014–4026, 2023.

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

[13] D. McDonald, C. Shalizi, and M. Schervish, "Estimating beta-mixing coefficients via histograms," *Electronic Journal of Statistics*, vol. 9, no. 2, pp. 2855–2883, 2015.

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

[15] 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*, vol. 29, no. 4, pp. 2439 – 2480, 2019.

[16] G. Wolfer and A. Kontorovich, "Estimating the mixing time of ergodic markov chains," in *Proceedings of the Thirty-Second Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 99, 25–28 Jun 2019, pp. 3120–3159.

[17] 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*, ser. Proceedings of Machine Learning Research, vol. 117, 2020, pp. 890–905.

[18] G. Wolfer and P. Alquier, "Optimistic estimation of convergence in markov chains with the average-mixing time," *arXiv preprint arXiv:2402.10506*, 2024.

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

[20] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications*, 4th ed. Springer, 2025.

[21] S. Janson, *Gaussian Hilbert Spaces*, ser. Cambridge Tracts in Mathematics. Cambridge University Press, 1997.

[22] S. Levental, "Uniform limit theorems for harris recurrent markov chains," *Probability theory and related fields*, vol. 80, no. 1, pp. 101–118, 1988.

[23] B. Yu, "Rates of Convergence for Empirical Processes of Stationary Mixing Sequences," *The Annals of Probability*, vol. 22, no. 1, pp. 94 – 116, 1994.

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

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

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

Steffen Grünewälder is a Reader in Statistics at the Department of Mathematics, University of York, UK. He received his B.Sc. and M.Sc. degrees from the University of Darmstadt, and his Ph.D. degree from TU-Berlin. His research interests include statistical learning theory, kernel methods and nonparametric statistics.

Azadeh Khaleghi is a professor of Statistics at École Nationale de la Statistique et de l'Administration Économique (ENSAE) - Centre de Recherche en Économie et Statistique (CREST). She received her B.Sc. and M.Sc. degrees from the University of Toronto, and her Ph.D. degree from INRIA - University of Lille I. Her research interests lie in statistical learning theory, nonparametric statistics, long-memory processes, and information theory.