Applied Probability Trust (17 January 2014)

UNIFORM CHERNOFF AND DVORETZKY-KIEFER-WOLFOWITZ-TYPE INEQUALITIES FOR MARKOV CHAINS AND RELATED PROCESSES

ARYEH KONTOROVICH,* *Ben-Gurion University of the Negev*

ROI WEISS,** *Ben-Gurion University of the Negev*

Abstract

We observe that the technique of Markov contraction can be used to establish measure concentration for a broad class of non-contracting chains. In particular, geometric ergodicity provides a simple and versatile framework. This leads to a short, elementary proof of a general concentration inequality for Markov and hidden Markov chains (HMM), which supercedes some of the known results and easily extends to other processes such as Markov trees. As applications, we give a Dvoretzky-Kiefer-Wolfowitz-type inequality and a uniform Chernoff bound. All of our bounds are dimension-free and hold for countably infinite state spaces.

Keywords: concentration of measure; Markov chain; HMM; Chernoff; Dvoretzky-Kiefer-Wolfowitz

2010 Mathematics Subject Classification: Primary 60E15

Secondary 60J10

1. Introduction

1.1. Background

The last two decades have seen a flurry of activity in concentration of measure for non-independent processes. A recent survey may be found in [19], with pointers to more specialized surveys therein. Rather than recapitulating these surveys here, we shall proceed directly to the relevant recent developments. Let $X_1, X_2, \ldots$ be a sequence of $\mathbb{N}$-valued random variables obeying some joint law (distribution). Using the shorthand

* Postal address: Department of Computer Science, Ben-Gurion University, Beer Sheva 84105, ISRAEL

$\mathcal{L}(X_j^n \,|\, X_1^i = x)$ to denote the law of $(X_j, \ldots, X_n)$ conditioned on $(X_1, \ldots, X_i) = x \in \mathbb{N}^i$, let us define, for $n \in \mathbb{N}$, $1 \leq i < j \leq n$, $y \in \mathbb{N}^{i-1}$ and $w, w' \in \mathbb{N}$,

$$\eta_{ij}(y, w, w') \quad = \quad \left\| \mathcal{L}(X_j^n \,|\, X_1^i = yw) - \mathcal{L}(X_j^n \,|\, X_1^i = yw') \right\|_{\text{TV}},$$

(where $\|\cdot\|_{\text{TV}} = \frac{1}{2}\|\cdot\|_1$ is the total variation norm) and

$$\bar{\eta}_{ij} \quad = \quad \sup_{y \in \mathbb{N}^{i-1}, w, w' \in \mathbb{N}} \eta_{ij}(y, w, w'). \tag{1}$$

The coefficients $\bar{\eta}_{ij}$, termed *η-mixing coefficients* in [21], play a central role in several recent concentration results. Define Δ to be the upper-triangular $n \times n$ matrix, with $\Delta_{ii} = 1$ and $\Delta_{ij} = \bar{\eta}_{ij}$ for $1 \leq i < j \leq n$.

In 2007, [7] and [21] independently proved that for any $f : \mathbb{N}^n \to \mathbb{R}$ with $\|f\|_{\text{Lip}} \leq 1$ with respect to the Hamming metric, we have

$$P\left(|f - \mathbb{E}f| > n\varepsilon\right) \quad \leq \quad 2\exp\left(-\frac{2n\varepsilon^2}{\min\left\{\|\Delta\|_2, \|\Delta\|_\infty\right\}^2}\right), \tag{2}$$

where $\|\Delta\|_p$ is the ℓ_p operator norm ([7] achieve the better constant in the exponent, given here). Earlier, Samson [34] had given a concentration result for convex ℓ_2-Lipschiz functions $f : [0,1]^n \to \mathbb{R}$, which likewise involved the coefficients $\bar{\eta}_{ij}$, and these are also implicit in Marton's earlier work [27, 28, 29].

In order to apply (2) in a Markov setting, one must upper-bound $\|\Delta\|_2$ or $\|\Delta\|_\infty$ for the Markov chain in question. The earliest such results relied on contraction. Let $p(\cdot \,|\, \cdot)$ be the transition kernel associated with a given Markov chain, and define the (Döblin) *contraction coefficient*

$$\kappa = \sup_{x, x' \in \mathbb{N}} \|p(\cdot \,|\, x) - p(\cdot \,|\, x')\|_{\text{TV}}. \tag{3}$$

It is shown in [21] and [34] that $\bar{\eta}_{ij} \leq \kappa^{j-i}$ and therefore $\|\Delta\|_\infty \leq (1-\kappa)^{-1}$; this implies the concentration bound

$$\mathbb{P}(|f - \mathbb{E}f| > n\varepsilon) \leq 2\exp(-2(1-\kappa)^2 n\varepsilon^2)$$

for 1-Lipschitz functions f, which Marton [26] had essentially obtained earlier by other means.

Meaning: if $x, y \in \mathbb{N}^n$ differ in only 1 coordinate then $|f(x) - f(y)| \leq 1$, see Section 2.7.

The contraction method was pushed further to obtain concentration results for hidden Markov chains [21], undirected Markov chains and Markov tree processes [19], but its applicability requires the rather stringent condition that $\kappa < 1$. Already in [27], Marton observed that a significantly weaker mixing condition suffices, and yields tighter and more informative bounds. Indeed, consider a Markov chain with stationary distribution π and conditional s^{th} step distribution $\mathcal{L}(X_s \,|\, X_1 = x)$, and define the "inverse mixing time"

$$\tau_s = \sup_{x \in \mathbb{N}} \left\| \mathcal{L}(X_s \,|\, X_1 = x) - \pi \right\|_{\text{TV}}. \tag{4}$$

A simple calculation (Lemma 2.3) shows that $\bar{\eta}_{ij} \leq 2\tau_{j-i}$, and thus

$$\left\| \Delta \right\|_\infty - 1 = \max_{1<i<n} \sum_{j=i+1}^{n} \bar{\eta}_{ij} \leq 2 \max_{1<i<n} \sum_{j=i+1}^{n} \tau_{j-i}.$$

A rich body of work deals with bounding τ_s via spectral [15], Poincaré [11], log-Sobolev [10] and Lyapunov [22] methods, among others (see the references in the works cited). From our perspective, the *geometric ergodicity* condition allows for the simplest exposition while sacrificing the least generality. A Markov chain is said to be geometrically ergodic with constants $1 \leq G < \infty$ and $0 \leq \theta < 1$ if

$$\tau_s \leq G\theta^{s-1}, \qquad s = 1, 2, \ldots. \tag{5}$$

Any finite ergodic Markov chain is geometrically ergodic, and the dependence of G, θ on various structural properties of the chain in question is the subject of a diverse and prolific literature (including the references above). We also stress that the geometric ergodicity assumption is largely dictated by expositional convenience, since any non-trivial bound on the inverse mixing time τ_s will yield straightforward analogues of our results.

In this paper, we explore some consequences of geometric ergodicity as pertaining to concentration and statistical inference for Markov and hidden Markov chains. We leverage two basic insights: (i) even though hidden Markov chains are a considerably richer class of processes than Markov chains (there exist HMMs not realizable by any finite-order Markov chain), for the purposes of measure concentration, the underlying

This terminology is non-standard.

Markov chain is all that matters and (ii) geometric ergodicity, while significantly more general than contractivity, yields essentially the same concentration bounds. Another advantage of our approach is its elementary nature: taking the bound in (2) as a given, nothing beyond basic linear algebra is used.

Given the recent interest in prediction and parameter inference for HMMs [3, 17, 31, 35, 20, 32], our result have potential to be applicable beyond the abstract setting studied here. Furthermore, since concentration results for Markov chains extend easily for other Markov-type processes (such as trees [19]), our results here should extend to those as well.

1.2. Main results

Concentration. Our first result is a concentration inequality for hidden Markov chains, which generalizes many of the previous such bounds. We will henceforth write "(G,θ)-geometrically ergodic" as shorthand for "geometrically ergodic with constants $1 \leq G < \infty$ and $0 \leq \theta < 1$". Hidden Markov chains and their associated notions of stationarity and geometric ergodicity are formally defined in Section 2.1.

Theorem 1.1. *Let $Y_1, Y_2, \ldots$ be a $\mathbb{N}$-valued hidden Markov chain whose underlying $\mathbb{N}$-valued Markov chain is (G,θ)-geometrically ergodic. Then, for any $n \in \mathbb{N}$ and $f : \mathbb{N}^n \to \mathbb{R}$ with $\|f\|_{\text{Lip}} \leq 1$ under the Hamming metric (see Section 2.7), we have*

$$\mathbb{P}\left(f(Y_1^n) - \mathbb{E}f(Y_1^n) > n\varepsilon\right) \quad \leq \quad \exp\left(-\frac{n(1-\theta)^2\varepsilon^2}{2G^2}\right),$$

with an identical bound for the other tail.

Although the result in Theorem 1.1 does not appear to have been published anywhere, it is a simple consequence of widely known facts (we give a proof in Section 2 for completeness). Our main contribution lies in the apparently novel applications.

DKW-type inequality. Let us recall the Dvoretzky-Kiefer-Wolfowitz inequality [14, 30], stated here for the discrete case. Suppose $X_1, X_2, \ldots$ are iid $\mathbb{N}$-valued random variables with common distribution function F, and define the empirical distribution function $\hat{F}_n$ induced by $(X_1, \ldots, X_n)$:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}, \qquad x \in \mathbb{N}.$$

The DKW inequality states that

$$\mathbb{P}\left(\sup_{x\in\mathbb{N}}\left|\hat{F}_n(x)-F(x)\right|>\varepsilon\right)\leq 2\exp(-2n\varepsilon^2),\qquad \varepsilon>0, n\in\mathbb{N}.$$

We present the following Markovian version of this inequality.

Theorem 1.2. *Let $Y_1, Y_2, \ldots$ be a stationary $\mathbb{N}$-valued (G,θ)-geometrically ergodic Markov or hidden Markov chain with stationary distribution $\rho\in\mathbb{R}^{\mathbb{N}}$. For $n\in\mathbb{N}$, define $\hat{\rho}^{(n)}\in\mathbb{R}^{\mathbb{N}}$ to be the empirical estimate of ρ:*

$$\hat{\rho}^{(n)}_y=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{Y_i=y\}},\qquad y\in\mathbb{N}. \tag{6}$$

Then

$$\mathbb{P}\left(\left\|\rho-\hat{\rho}^{(n)}\right\|_\infty>\sqrt{\frac{1+2G\theta}{n(1-\theta)}}+\varepsilon\right)\leq\exp\left(-\frac{n(1-\theta)^2\varepsilon^2}{2G^2}\right),\qquad n\in\mathbb{N},\varepsilon>0.$$

As we show in Section 2.6, the assumption that the chain starts in the stationary distribution is not at all restrictive. Note that a naive application of Theorem 1.1 to each $\hat{\rho}^{(n)}_y$ individually, combined with the union bound, would yield

$$\mathbb{P}\left(\left\|\rho-\hat{\rho}^{(n)}\right\|_\infty>\varepsilon\right)\leq 2\left\|\rho\right\|_0\exp\left(-\frac{n(1-\theta)^2\varepsilon^2}{2G^2}\right), \tag{7}$$

where $\|\rho\|_0$ is the number of non-zero entries in ρ. The bound in (7) is considerably weaker than the one in Theorem 1.2 and in particular is vacuous for ρ with infinite support.

Uniform Chernoff bound. Let $Y_1, Y_2, \ldots$ be a stationary $\mathbb{N}$-valued (G,θ)-geometrically ergodic Markov or hidden Markov chain as above, and consider the occupation frequency:

$$\hat{\rho}^{(n)}(E)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{Y_i\in E\}},\qquad E\subseteq\mathbb{N}.$$

A naive application of Theorem 1.1 might yield a deviation bound along the lines of

$$\mathbb{P}\left(\left|\rho(E)-\hat{\rho}^{(n)}(E)\right|>\varepsilon\right)\leq 2|E|\exp\left(-\frac{n(1-\theta)^2\varepsilon^2}{2|E|^2G^2}\right),$$

where $|E|$ is the cardinality of E and ρ is the stationary distribution as above. We will give a much stronger bound, that is not only independent of E but is actually uniform over all $E\subseteq\mathbb{N}$.

Theorem 1.3. *For stationary (G,θ)-geometrically ergodic (hidden) Markov chains, define*

$$\Lambda_n(\rho) = \gamma_n(G,\theta)\sum_{\rho_y \geq 1/n}\sqrt{\rho_y} + \min\left\{\gamma_n(G,\theta)\sum_{\rho_y<1/n}\sqrt{\rho_y}\ , \sum_{\rho_y<1/n}\rho_y\right\}, \qquad n\in\mathbb{N},$$

where

$$\gamma_n(G,\theta) = \frac{1}{2}\sqrt{\frac{1+2G\theta}{n(1-\theta)}}.$$

Then:

(a) for all distributions $\rho \in \mathbb{R}^{\mathbb{N}}$,

$$\lim_{n\to\infty}\Lambda_n(\rho) = 0,$$

(b)

$$\mathbb{P}\left(\sup_{E\subseteq\mathbb{N}}\left|\rho(E)-\hat{\rho}^{(n)}(E)\right| > \Lambda_n(\rho)+\varepsilon\right) \leq \exp\left(-\frac{n(1-\theta)^2\varepsilon^2}{2G^2}\right).$$

We remark that although $\Lambda_n(\rho) \underset{n\to\infty}{\longrightarrow} 0$ for all stationary distributions ρ, the rate of decay is ρ-dependent and may be arbitrarily slow for heavy-tailed distributions (cf. [4, Lemma 8]). For ρ satisfying $\sum_{y\in\mathbb{N}}\sqrt{\rho_y} < \infty$, the bound in Theorem 1.3 may be somewhat simplified via

$$\Lambda_n(\rho) \leq \gamma_n(G,\theta)\sum_{y\in\mathbb{N}}\sqrt{\rho_y}.$$

As with Theorem 1.2, the stationarity assumption $p_1 = \pi$ is not very restrictive (see Section 2.6).

1.3. Related work

In parallel to the work on concentration of measure results for Markov chains [1, 2, 8, 21, 26, 34], grew a body of Chernoff-type bounds for these processes. The papers [12, 13, 16, 18, 24] played a founding role, and various extensions and refinements followed [23, 36]. In a remarkable recent development [9], optimal Chernoff-Hoeffding bounds are obtained based on the mixing time at a constant threshold.

Concentration of Lipschitz functions of mixing sequences, with applications to the Kolmogorov-Smirnov statistic, were considered in [33]. The paper [5] examines the concentration of empirical distributions for non-independent sequences satisfying Poincaré or log-Sobolev inequalities.

2. Methods and proofs

2.1. Preliminaries

For readability, we will sometimes write the matrix entry $A_{x,y}$ as $A(x \,|\, y)$. We will use the terms *hidden Markov chain* and HMM interchangeably.

Markov chains. We will represent Markov kernels by column-stochastic $\mathbb{N} \times \mathbb{N}$ matrices denoted by the letter A. Thus, a Markov chain with transition kernel A and initial distribution p_1 induces the following distribution on $\mathbb{N}^n$:

$$\mathcal{L}(X_1, \ldots, X_n) = p_1(X_1) \prod_{i=1}^{n-1} A(X_{i+1} \,|\, X_i). \tag{8}$$

Hidden Markov chain. A hidden Markov chain (also known as hidden Markov model [HMM]) is specified by the triple (p_1, A, B), where (p_1, A) are the Markov chain parameters as above and B is an $\mathbb{N} \times \mathbb{N}$ column-stochastic matrix of *emission probabilities*. This HMM induces a distribution on $\mathbb{N}^n$ as follows. Let $X \in \mathbb{N}^n$ be distributed according to (8) and define the conditional distribution $\mathcal{L}(\cdot \,|\, X)$ over $Y \in \mathbb{N}^n$:

$$\mathcal{L}(Y \,|\, X) = \prod_{i=1}^{n} B(Y_i \,|\, X_i).$$

It follows that

$$\mathcal{L}(Y) = \sum_{x \in \mathbb{N}^n} \mathbb{P}(X = x) \mathcal{L}(Y \,|\, X = x).$$

We will refer to Y as a *hidden Markov* chain and to X as its *underlying* Markov chain.

Stationary distributions and chains. The stationary distribution $\pi \in \mathbb{R}^{\mathbb{N}}$ of the Markov chain with transition kernel A is the unique stochastic vector satisfying $A\pi = \pi$. The Markov chain induced by (p_1, A) is said to be stationary if $p_1 = \pi$. It is well-known that, for ergodic Markov chains,

$$\pi = \lim_{n \to \infty} \mathcal{L}(X_n) = \lim_{n \to \infty} \mathbb{E}\hat{\pi}^{(n)},$$

where

$$\hat{\pi}_x^{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i = x\}}, \qquad x \in \mathbb{N}.$$

In the geometrically ergodic case, observing that $\mathbb{E}\hat{\pi}^{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(X_i)$, we have

$$\begin{aligned}
\left\|\mathbb{E}\hat{\pi}^{(n)} - \pi\right\|_{\text{TV}} &= \left\|\frac{1}{n}\sum_{i=1}^{n}(\mathcal{L}(X_i) - \pi)\right\|_{\text{TV}} \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left\|\mathcal{L}(X_i) - \pi\right\|_{\text{TV}} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{x\in\mathbb{N}}\mathcal{L}(X_i \,|\, X_1 = x)p_1(x) - \pi\right\|_{\text{TV}} \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathbb{N}}p_1(x)\left\|\mathcal{L}(X_i \,|\, X_1 = x) - \pi\right\|_{\text{TV}} \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathbb{N}}p_1(x)G\theta^{i-1} = \frac{G}{(1-\theta)n}.
\end{aligned}$$

For a hidden Markov chain, we define the stationary distribution $\rho = B\pi$, and observe that

$$\rho = \lim_{n\to\infty}\mathcal{L}(Y_n) = \lim_{n\to\infty}\mathbb{E}\hat{\rho}^{(n)},$$

where $\hat{\rho}^{(n)}$ is defined in (6). Since $\hat{\rho}^{(n)}$ is distributed as $B\hat{\pi}^{(n)}$, we have

$$\left\|\mathbb{E}\hat{\rho}^{(n)} - \rho\right\|_{\text{TV}} \leq \left\|\mathbb{E}\hat{\pi}^{(n)} - \pi\right\|_{\text{TV}} \leq \frac{G}{(1-\theta)n}. \tag{9}$$

The bound in (9) suggests that, at least to some degree, the statistical behavior of an HMM is controlled by its underlying Markov chain. We expand upon this observation further:

Lemma 2.1. *Let X and X' be two Markov chains induced by (ξ, A) and (ξ', A'), respectively. For a given emission matrix B, let Y and Y' be the hidden Markov chains induced by (ξ, A, B) and (ξ', A', B). Then*

$$\left\|\mathcal{L}(Y_{i\in I}) - \mathcal{L}(Y'_{i\in I})\right\|_{\text{TV}} \leq \left\|\mathcal{L}(X_{i\in I}) - \mathcal{L}(X'_{i\in I})\right\|_{\text{TV}}, \qquad I \subseteq \{1,\ldots,n\}, n \in \mathbb{N}.$$

Proof. Immediate from Jensen's inequality, since hidden Markov chains are convex mixtures of Markov chains.

The proofs of Theorems 1.2 and 1.3 will require bounds on $\left\|\hat{\rho}^{(n)} - \rho\right\|$, but unlike in (9), the expectation is on the outside of the norm.

2.2. Markov contraction

Let us recast the contraction coefficient defined in (3) in the language of Markov kernels:

$$\kappa = \sup_{x,x' \in \mathbb{N}} \|A(\cdot \,|\, x) - A(\cdot \,|\, x')\|_{\mathrm{TV}} .$$

The term "contraction" is justified by the following simple fact [6, 21]:

Lemma 2.2. (Markov, 1906 [25].) *For any two stochastic vectors $\xi, \psi \in \mathbb{R}^{\mathbb{N}}$, we have*

$$\|A(\xi - \psi)\|_{\mathrm{TV}} \leq \kappa \, \|\xi - \psi\|_{\mathrm{TV}} .$$

Our principal application of this result will be in the context of geometrically ergodic Markov kernels.

Corollary 2.1. *Let A be a (G, θ)-geometrically ergodic Markov kernel. Then for all $n \in \mathbb{N}$, the n-step kernel A^n has contraction coefficient $\kappa \leq 2G\theta^n$.*

Proof. Let π be the stationary distribution of A and $\xi, \psi \in \mathbb{R}^{\mathbb{N}}$ two point masses. Then

$$\begin{aligned} \|A^n \xi - A^n \psi\|_{\mathrm{TV}} &\leq \|A^n \xi - \pi\|_{\mathrm{TV}} + \|A^n \psi - \pi\|_{\mathrm{TV}} \\ &\leq 2\tau_{n+1} \leq 2G\theta^n. \end{aligned}$$

2.3. Proof of main inequality

In this section, we prove Theorem 1.1. The first order of business is to bound the η-mixing coefficient by the inverse mixing time, and hence in terms of G and θ.

Lemma 2.3. *Let Y be a (G, θ)-geometrically ergodic hidden Markov chain and let $\bar{\eta}_{ij}$ and τ_s be as defined in (1) and (4), respectively. Then*

$$\bar{\eta}_{ij} \leq 2\tau_{j-i+1} \leq 2G\theta^{j-i}, \qquad n \in \mathbb{N}, 1 \leq i < j \leq n.$$

Proof. Let X be the Markov chain underlying Y and endow $\bar{\eta}_{ij}(X)$, $\bar{\eta}_{ij}(Y)$ with the obvious meaning. Then [21, Theorem 7.1] shows that

$$\bar{\eta}_{ij}(Y) \leq \bar{\eta}_{ij}(X).$$

Next, Remark 4 and the Theorem preceding it in [19] show that

$$\bar{\eta}_{ij}(X) \leq \kappa(A^{j-i})$$

where $\kappa(A^{j-i})$ is the contraction coefficient of the $(j-i)$-step Markov kernel of X. Finally, Corollary 2.1 yields

$$\kappa(A^{j-i}) \leq 2\tau_{j-i+1} \leq 2G\theta^{j-i}.$$

Proof of Theorem 1.1.. By (2), it suffices to upper-bound

$$\|\Delta\|_\infty = 1 + \max_{1<i<n} \sum_{j=i+1}^{n} \bar{\eta}_{ij}.$$

Applying Lemma 2.3, we get

$$\begin{aligned} \max_{1<i<n} \sum_{j=i+1}^{n} \bar{\eta}_{ij} &\leq 2G \max_{1<i<n} \sum_{j=i+1}^{n} \theta^{j-i} \\ &\leq 2G \sum_{k=1}^{\infty} \theta^k. \end{aligned}$$

Since $G \geq 1$ by assumption, we have

$$\begin{aligned} 1 + 2G \sum_{k=1}^{\infty} \theta^k &\leq 2G \sum_{k=0}^{\infty} \theta^k \\ &\leq \frac{2G}{1-\theta}. \end{aligned}$$

2.4. Proof of the DKW-type inequality

In this section, we prove Theorem 1.2. Let $Y_1, Y_2, \ldots$ be a stationary (G,θ)-geometrically ergodic hidden Markov chain with stationary distribution ρ, and define the $\{0,1\}$-indicator variables

$$\xi_i^{(y)} = \mathbb{1}_{\{Y_i = y\}}, \qquad i, y \in \mathbb{N}. \tag{10}$$

Then $\hat{\rho}$, defined in (6), is given by $\hat{\rho}_y = \frac{1}{n}\sum_{i=1}^{n} \xi_i^{(y)}$, where we have dropped the superscript (n) from $\hat{\rho}$ for readability. Observing that the map $(Y_1, \ldots, Y_n) \mapsto n\,\|\rho - \hat{\rho}\|_\infty$ is 1-Lipschitz under the Hamming metric (Lemma 2.7), we apply Theorem 1.1:

$$\mathbb{P}(\|\rho - \hat{\rho}\|_\infty > \mathbb{E}\,\|\rho - \hat{\rho}\|_\infty + \varepsilon) \leq \exp\left(-\frac{n(1-\theta)^2\varepsilon^2}{2G^2}\right).$$

Hence, it remains to bound $\mathbb{E}\,\|\rho - \hat{\rho}\|_\infty$.

Lemma 2.4.

$$\mathbb{E}\left\|\rho-\hat{\rho}\right\|_{\infty} \quad \leq \quad \sqrt{\frac{1+2G\theta}{n(1-\theta)}}.$$

Remark. This estimate is nearly optimal: in the case where Y_i are iid (i.e., $\theta = 0$) Bernoulli variables with parameter p, we have [4, Theorem 1]

$$\sqrt{\frac{p(1-p)}{2n}} \leq \mathbb{E}\left\|\rho-\hat{\rho}\right\|_{\infty} \leq \sqrt{\frac{p(1-p)}{n}}, \qquad n \geq 2,\ p \in [1/n, 1-1/n].$$

Proof. Jensen's inequality yields

$$\begin{aligned}
\left(\mathbb{E}\left\|\rho-\hat{\rho}\right\|_{\infty}\right)^2 &\leq \mathbb{E}\left[\left\|\rho-\hat{\rho}\right\|_{\infty}^2\right] \\
&\leq \mathbb{E}\left[\sum_{y\in\mathbb{N}}\left|\rho_y-\hat{\rho}_y\right|^2\right] \qquad (11)\\
&= \sum_{y\in\mathbb{N}}\mathbb{E}\left(\rho_y-\hat{\rho}_y\right)^2 = \sum_{y\in\mathbb{N}}\mathrm{Var}[\hat{\rho}_y].
\end{aligned}$$

Putting $S_n^{(y)} = \sum_{i=1}^n \xi_i^{(y)}$, we have

$$n^2\mathrm{Var}[\hat{\rho}_y] = \mathbb{E}\Big(S_n^{(y)}\Big)^2 - \Big(\mathbb{E}S_n^{(y)}\Big)^2 \qquad (12)$$

and

$$\mathbb{E}S_n^{(y)} = n\rho_y. \qquad (13)$$

To bound $\mathbb{E}\Big(S_n^{(y)}\Big)^2$, we compute

$$\begin{aligned}
\mathbb{E}\Big(S_n^{(y)}\Big)^2 &= \mathbb{E}\left[\sum_{1\leq i,j\leq n}\xi_i^{(y)}\xi_j^{(y)}\right] \\
&= \sum_{i=1}^n \mathbb{E}\Big(\xi_i^{(y)}\Big)^2 + 2\sum_{1\leq i<j\leq n}\mathbb{E}\left[\xi_i^{(y)}\xi_j^{(y)}\right] \\
&= n\rho_y + 2\sum_{1\leq i<j\leq n}\mathbb{E}\left[\xi_i^{(y)}\xi_j^{(y)}\right], \qquad (14)
\end{aligned}$$

where the last identity holds since $\xi_i^{(y)} \in \{0,1\}$. It now remains to estimate $\mathbb{E}\left[\xi_i^{(y)}\xi_j^{(y)}\right]$. To this end, we claim that

$$\left\|\mathcal{L}(Y_i \,|\, Y_1 = y) - \rho\right\|_{\infty} \leq G\theta^{i-1}, \qquad i, y \in \mathbb{N}.$$

Indeed, denoting the parameters of Y by (π, A, B) and letting X be the underlying Markov chain, we have

$$\begin{aligned}
\|\mathcal{L}(Y_i \,|\, Y_1 = y_1) - \rho\|_\infty &\leq \|\mathcal{L}(Y_i \,|\, Y_1 = y_1) - \rho\|_{\text{TV}} \\
&= \frac{1}{2} \sum_{y_i \in \mathbb{N}} |\mathbb{P}(Y_i = y_i \,|\, Y_1 = y_1) - \rho_{y_i}| \\
&= \frac{1}{2} \sum_{y_i \in \mathbb{N}} \left| \sum_{x_i \in \mathbb{N}} B_{y_i, x_i} \left(\mathbb{P}(X_i = x_i \,|\, Y_1 = y_1) - \pi_{x_i} \right) \right| \\
&\leq \frac{1}{2} \sum_{y_i \in \mathbb{N}} \sum_{x_i \in \mathbb{N}} B_{y_i, x_i} |\mathbb{P}(X_i = x_i \,|\, Y_1 = y_1) - \pi_{x_i}| \\
&= \frac{1}{2} \sum_{x_i \in \mathbb{N}} |\mathbb{P}(X_i = x_i \,|\, Y_1 = y_1) - \pi_{x_i}| \\
&= \left\| \sum_{x_1 \in \mathbb{N}} \mathcal{L}(X_i \,|\, X_1 = x_1) \mathbb{P}(X_1 = x_1 \,|\, Y_1 = y_1) - \pi \right\|_{\text{TV}} \\
&\leq \sup_{x_1 \in \mathbb{N}} \|\mathcal{L}(X_i \,|\, X_1 = x_1) - \pi\|_{\text{TV}} \leq G\theta^{i-1}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[\xi_i^{(y)} \xi_j^{(y)}\right] &= \mathbb{P}(Y_i = y, Y_j = y) \\
&= \mathbb{P}(Y_1 = y, Y_{j-i+1} = y) \\
&= \mathbb{P}(Y_1 = y) \mathbb{P}(Y_{j-i+1} = y \,|\, Y_1 = y) \\
&\leq \rho_y(\rho_y + G\theta^{j-i}),
\end{aligned}$$

and therefore

$$\begin{aligned}
\sum_{1 \leq i < j \leq n} \mathbb{E}\left[\xi_i^{(y)} \xi_j^{(y)}\right] &= \sum_{k=1}^{n-1} (n-k) \mathbb{P}(Y_1 = y) \mathbb{P}(Y_{k+1} = y \,|\, Y_1 = y) \\
&\leq \sum_{k=1}^{n-1} (n-k) \rho_y(\rho_y + G\theta^k) \\
&= \frac{n(n-1)}{2} \rho_y^2 + \frac{G\theta}{1-\theta} \left(n - \frac{1-\theta^n}{1-\theta} \right) \rho_y \\
&\leq \frac{n(n-1)}{2} \rho_y^2 + n \frac{G\theta}{1-\theta} \rho_y. \qquad (15)
\end{aligned}$$

Combining (12), (13), (14), and (15), we have

$$\begin{aligned}\mathrm{Var}[\hat{\rho}_y] &\leq \frac{1}{n^2}\left(n\rho_y + n(n-1)\rho_y^2 + 2n\frac{G\theta}{1-\theta}\rho_y - n^2\rho_y^2\right)\\ &= \frac{\rho_y}{n}\left(1-\rho_y+\frac{2G\theta}{1-\theta}\right)\\ &\leq \rho_y\frac{1+2G\theta}{n(1-\theta)}.\end{aligned}$$

Since $\sum_{y\in\mathbb{N}}\rho_y = 1$, the claim follows from (11).

Remark. Note that in the process of proving a deviation estimate on $\|\rho-\hat{\rho}\|_\infty$, we have actually proven a stronger one — namely, for the ℓ_2 norm.

2.5. Proof of the uniform Chernoff bound

In this section, we prove Theorem 1.3. As before, $Y_1, Y_2, \ldots$ is a stationary (G,θ)-geometrically ergodic hidden Markov chain with stationary distribution ρ. Since by Lemma 2.7 the map $(Y_1,\ldots,Y_n) \mapsto n\|\rho-\hat{\rho}\|_{\mathrm{TV}}$ is 1-Lipschitz under the Hamming metric, Theorem 1.1 applies:

$$\mathbb{P}(\|\rho-\hat{\rho}\|_{\mathrm{TV}} > \mathbb{E}\|\rho-\hat{\rho}\|_{\mathrm{TV}} + \varepsilon) \leq \exp\left(-\frac{n(1-\theta)^2\varepsilon^2}{2G^2}\right). \tag{16}$$

As before, the crux of the matter is to bound $\mathbb{E}\|\rho-\hat{\rho}\|_{\mathrm{TV}}$. Recall the definition of Λ_n from the statement of Theorem 1.3.

Lemma 2.5.

$$\mathbb{E}\|\rho-\hat{\rho}\|_{\mathrm{TV}} \leq \Lambda_n.$$

Remark. This bound is nearly optimal: when the Y_i are iid, we have [4, Proposition 3]

$$\mathbb{E}\|\rho-\hat{\rho}\|_{\mathrm{TV}} \geq \frac{1}{4}\Lambda_n - \frac{1}{8\sqrt{n}}, \qquad n\geq 2,\ p\in[1/n, 1-1/n].$$

Proof. We proceed by breaking up the expectation into two terms,

$$\mathbb{E}\|\rho-\hat{\rho}\|_{\mathrm{TV}} = \tfrac{1}{2}\sum_{y:\rho_y<1/n}\mathbb{E}|\rho_y-\hat{\rho}_y| + \tfrac{1}{2}\sum_{y:\rho_y\geq 1/n}\mathbb{E}|\rho_y-\hat{\rho}_y|, \tag{17}$$

and bounding each term separately. To bound the second term, we note, as in the proof of Lemma 2.4, that

$$\mathbb{E}|\rho_y-\hat{\rho}_y| \leq \sqrt{\mathrm{Var}[\hat{\rho}_y]} \leq \sqrt{\rho_y\frac{1+2G\theta}{n(1-\theta)}}, \qquad y\in\mathbb{N}. \tag{18}$$

To bound the first term, we recall the indicator variables $\xi_i^{(y)}$ defined in (10) and observe that

$$\begin{aligned} n\mathbb{E}\left|\rho_y - \hat{\rho}_y\right| &= \mathbb{E}\left|\sum_{i=1}^{n} \xi_i^{(y)} - n\rho_y\right| \\ &\leq n\mathbb{E}\left|\xi_i^{(y)} - \rho_y\right| \\ &= 2n\rho_y(1-\rho_y) \leq 2n\rho_y, \end{aligned}$$

where stationarity was used in the last line of the derivation.

Combining the last display with (17) and (18) yields the claim.

Proof of Theorem 1.3.. (a) Since obviously

$$\sum_{\rho_y < 1/n} \rho_y \quad \underset{n\to\infty}{\longrightarrow} \quad 0,$$

it suffices to show that

$$\frac{1}{\sqrt{n}} \sum_{\rho_y \geq 1/n} \sqrt{\rho_y} \quad \underset{n\to\infty}{\longrightarrow} \quad 0. \tag{19}$$

The latter was proved in [4, Lemma 7], but we will present a simpler proof here. Assume without loss of generality that $\rho_1 \geq \rho_2 \geq \ldots$, pick an abitrary $\varepsilon > 0$, and let $N \in \mathbb{N}$ be large enough so that $\sum_{j \geq N} \rho_j < \varepsilon$. Then

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{\rho_j \geq 1/n} \sqrt{\rho_j} &\leq \frac{1}{\sqrt{n}} \sum_{j \leq N} \sqrt{\rho_j} + \frac{1}{\sqrt{n}} \sum_{j > N, \rho_j \geq 1/n} \sqrt{\rho_j} \\ &\leq \sqrt{\frac{N}{n}} + \frac{1}{\sqrt{n}} \sqrt{\sum_{\rho_j \geq 1/n} 1} \sqrt{\sum_{j>N} \rho_j} \\ &\leq \sqrt{\frac{N}{n}} + \sqrt{\varepsilon}, \end{aligned}$$

since there can be at most n terms with $\rho_j \geq 1/n$.

(b) The claim follows from (16) and the fact that for any two distributions $\phi, \psi \in \mathbb{R}^{\mathbb{N}}$,

$$\|\phi - \psi\|_{\text{TV}} = \sup_{E \subseteq \mathbb{N}} \left|\phi(E) - \psi(E)\right|.$$

This elegant proof is due to Asaf Shachar. Andrew Barron points out that (19) may be easily derived from Lebesgue's dominated convergence theorem.

2.6. The stationarity assumption

To state the bounds in Theorems 1.2 and 1.3 more cleanly, we had assumed that the Markov and hidden Markov chains in question are stationary — i.e., that the initial distribution p_1 is identical to the stationary one π. In this section we show (Corollary 2.2) that for strongly mixing chains, the stationarity assumption may be relaxed, at the cost of additional terms in the deviation bounds.

Let $Y = (Y_1, \ldots, Y_n)$ be a (G, θ)-geometrically ergodic hidden Markov chain with parameters (π', A, B), where $\pi' \in \mathbb{R}^{\mathbb{N}}$ is some stochastic vector. A simple dimension-free bound on the statistical distance between Y and its stationary version is available:

Theorem 2.1. *Let $Y' = (Y'_1, \ldots, Y'_n)$ be the stationary version of Y — i.e., an HMM with parameters (π, A, B), where π is the stationary distribution of the kernel A. Then*

$$\left\|\mathcal{L}(Y) - \mathcal{L}(Y')\right\|_{\mathrm{TV}} \leq \left\|\pi - \pi'\right\|_{\mathrm{TV}}.$$

First, we prove an analogous result for Markov chains.

Lemma 2.6. *Let A be Markov kernel and $\xi, \xi' \in \mathbb{R}^{\mathbb{N}}$ two arbitrary stochastic vectors. Let $X = (X_1, \ldots, X_n)$ and $X' = (X'_1, \ldots, X'_n)$ be the Markov chains induced by (ξ, A) and (ξ', A), respectively. Then*

$$\left\|\mathcal{L}(X) - \mathcal{L}(X')\right\|_{\mathrm{TV}} = \left\|\xi - \xi'\right\|_{\mathrm{TV}}.$$

Proof.

$$\begin{aligned}
\left\|\mathcal{L}(X) - \mathcal{L}(X')\right\|_{\mathrm{TV}} &= \frac{1}{2} \sum_{x \in \mathbb{N}^n} \left| \left(\xi_{x_1} - \xi'_{x_1}\right) A_{x_2,x_1} \ldots A_{x_n,x_{n-1}} \right| \\
&= \frac{1}{2} \sum_{x \in \mathbb{N}^n} A_{x_2,x_1} \ldots A_{x_n,x_{n-1}} \left| \xi_{x_1} - \xi'_{x_1} \right| \\
&= \frac{1}{2} \sum_{x_1 \in \mathbb{N}} \left| \xi_{x_1} - \xi'_{x_1} \right| = \left\|\xi - \xi'\right\|_{\mathrm{TV}}.
\end{aligned}$$

Proof of Theorem 2.1. Lemma 2.1 lets us restrict our attention to the underlying Markov chains X and X', respectively:

$$\begin{aligned}
\left\|\mathcal{L}(Y_{1 \leq i \leq n}) - \mathcal{L}(Y'_{1 \leq i \leq n})\right\|_{\mathrm{TV}} &\leq \left\|\mathcal{L}(X_{1 \leq i \leq n}) - \mathcal{L}(X'_{1 \leq i \leq n})\right\|_{\mathrm{TV}} \\
&= \left\|\mathcal{L}(X_1) - \mathcal{L}(X'_1)\right\|_{\mathrm{TV}} = \left\|\pi - \pi'\right\|_{\mathrm{TV}},
\end{aligned}$$

where the first identity follows from Lemma 2.6.

Corollary 2.2. *Let $Y_1, Y_2, \ldots$ be a (not necessarily stationary) $\mathbb{N}$-valued (G,θ)-geometrically ergodic hidden Markov chain with stationary distribution $\rho = B\pi$ and initial distribution $\rho' = B\pi$. Then the deviation bounds stated in Theorems 1.2 and 1.3 hold with an additive correction of $\|\pi - \pi'\|_{\mathrm{TV}}$ on the right-hand side.*

2.7. Auxiliary lemma

The Hamming metric on $\mathbb{N}^n$ is defined by $d(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^n \mathbb{1}_{\{x_i \neq y_i\}}$ for $\boldsymbol{x},\boldsymbol{y} \in \mathbb{N}^n$.

Lemma 2.7. *Suppose $n \in \mathbb{N}$ and $\boldsymbol{p} \in \mathbb{R}^{\mathbb{N}}$ is a distribution. Define the functions $g, h : \mathbb{N}^n \to \mathbb{R}$ by*

$$\begin{aligned} g(\boldsymbol{x}) &= \sup_{j\in\mathbb{N}} \left| np_j - \sum_{i=1}^n \mathbb{1}_{\{x_i=j\}} \right|, \qquad \boldsymbol{x} \in \mathbb{N}^n, \\ h(\boldsymbol{x}) &= \sum_{j\in\mathbb{N}} \left| np_j - \sum_{i=1}^n \mathbb{1}_{\{x_i=j\}} \right|, \qquad \boldsymbol{x} \in \mathbb{N}^n. \end{aligned}$$

Then $\|g\|_{\mathrm{Lip}} \leq 1$ and $\|h\|_{\mathrm{Lip}} \leq 2$ with respect to the Hamming metric:

$$\begin{aligned} |g(\boldsymbol{x}) - g(\boldsymbol{y})| &\leq d(\boldsymbol{x},\boldsymbol{y}), \\ |h(\boldsymbol{x}) - h(\boldsymbol{y})| &\leq 2d(\boldsymbol{x},\boldsymbol{y}) \end{aligned}$$

for all $\boldsymbol{x},\boldsymbol{y} \in \mathbb{N}^n$.

Proof. We only prove the claim for h (the proof for g is analogous). Let the function $\hat{n}_j : \mathbb{N}^n \to \mathbb{N}$ count the number of times j appears in x; formally, $\hat{n}_j(x) = \sum_{i=1}^n \mathbb{1}_{\{x_i=j\}}$. Now suppose $\boldsymbol{x},\boldsymbol{y} \in \mathbb{N}^n$ differ only in coordinate k, with $x_k = a$ and $y_k = b$. Then

$$\begin{aligned} h(\boldsymbol{x}) - h(\boldsymbol{y}) &= \sum_{j\in\mathbb{N}} |np_j - \hat{n}_j(\boldsymbol{x})| - \sum_{j\in\mathbb{N}} |np_j - \hat{n}_j(\boldsymbol{y})| \\ &= (|np_a - \hat{n}_a(\boldsymbol{x})| + |np_b - \hat{n}_b(\boldsymbol{x})|) - (|np_a - \hat{n}_a(\boldsymbol{y})| + |np_b - \hat{n}_b(\boldsymbol{y})|) \\ &= (|np_a - \hat{n}_a(\boldsymbol{x})| + |np_b - \hat{n}_b(\boldsymbol{x})|) - (|np_a - (\hat{n}_a(\boldsymbol{x}) - 1)| + |np_b - (\hat{n}_b(\boldsymbol{x}) + 1)|) \\ &\leq \big||np_a - \hat{n}_a(\boldsymbol{x})| - |np_a - (\hat{n}_a(\boldsymbol{x}) - 1)|\big| + \big||np_b - \hat{n}_b(\boldsymbol{x})| - |np_b - (\hat{n}_b(\boldsymbol{x}) + 1)|\big| \\ &\leq 2. \end{aligned}$$

Acknowledgments

We thank the anonymous referee for carefully reading the manuscript and offering helpful suggestions.

References

[1] Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13,** no. 34, 1000–1034.

[2] Adamczak, R. and Bednorz, W. (2012). Exponential concentration inequalities for additive functionals of Markov chains (arxiv:1201.3569v1).

[3] Anandkumar, A., Hsu, D. and Kakade, S. M. (2012). A method of moments for mixture models and hidden Markov models. In *COLT*.

[4] Berend, D. and Kontorovich, A. (2013). A sharp estimate of the binomial mean absolute deviation with applications. *Statistics & Probability Letters* **83,** 1254–1259.

[5] Bobkov, S. G. and Götze, F. (2010). Concentration of empirical distribution functions with applications to non-i.i.d. models. *Bernoulli* **16,** 1385–1414.

[6] Brémaud, P. (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer-Verlag.

[7] Chazottes, J.-R., Collet, P., Külske, C. and Redig, F. (2007). Concentration inequalities for random fields via coupling. *Probability Theory and Related Fields* **137,** 201–225.

[8] Chazottes, J.-R. and Redig, F. (2009). Concentration inequalities for Markov processes via coupling. *Electron. J. Probab.* **14,** no. 40, 1162–1180.

[9] Chung, K.-M., Lam, H., Liu, Z. and Mitzenmacher, M. (2012). Chernoff-hoeffding bounds for Markov chains: Generalized and simplified. In *STACS*. pp. 124–135.

[10] Diaconis, P. and Saloff-Coste, L. (1996). Logarithmic Sobolev inequalities for finite Markov chains. *Ann. Appl. Probab.* **6,** 695–750.

[11] Diaconis, P. and Saloff-Coste, L. (1996). Nash inequalities for finite Markov chains. *Journal of Theoretical Probability* **9,** 459–510.

[12] Dinwoodie, I. H. (1995). A probability inequality for the occupation measure of a reversible Markov chain. *Ann. Appl. Probab.* **5,** 37–43.

[13] Dinwoodie, I. H. (1998). Expectations for nonreversible Markov chains. *Journal of Mathematical Analysis and Applications* **220,** 585 – 596.

[14] Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **27,** 642–669.

[15] Fill, J. A. (1991). Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *Ann. Appl. Probab.* **1,** 62–87.

[16] Gillman, D. (1998). A Chernoff bound for random walks on expander graphs. *SIAM J. Comput.* **27,** 1203–1220.

[17] Hsu, D., Kakade, S. M. and Zhang, T. (2009). A spectral algorithm for learning hidden Markov models. In *COLT*.

[18] Kahale, N. (1997). Large deviation bounds for Markov chains. *Combinatorics, Probability & Computing* **6,** 465–474.

[19] Kontorovich, A. (2012). Obtaining measure concentration from Markov contraction. *Markov Processes and Related Fields* **4,** 613–638.

[20] Kontorovich, A., Nadler, B. and Weiss, R. (2013). On learning parametric-output hmms. In *ICML (3)*. pp. 702–710.

[21] Kontorovich, L. A. and Ramanan, K. (2008). Concentration Inequalities for Dependent Random Variables via the Martingale Method. *Ann. Probab.* **36,** 2126–2158.

[22] Kontoyiannis, I. and Meyn, S. (2012). Geometric ergodicity and the spectral gap of non-reversible Markov chains. *Probability Theory and Related Fields* **154,** 327–339.

[23] León, C. A. and Perron, F. (2004). Optimal Hoeffding bounds for discrete reversible Markov chains. *Ann. Appl. Probab.* **14,** 958–970.

[24] Lezaud, P. (1998). Chernoff-type bound for finite Markov chains. *Ann. Appl. Probab.* **8,** 849–867.

[25] Markov, A. A. (1906). Extension of the law of large numbers to dependent quantities. *Izvestiia Fiz.-Matem. Obsch. Kazan Univ.* **15,** 135–156.

[26] Marton, K. (1996). Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration. *Ann. Probab.* **24,** 857–866.

[27] Marton, K. (1998). Measure concentration for a class of random processes. *Probability Theory and Related Fields* **110,** 427–439.

[28] Marton, K. (2003). Measure concentration and strong mixing. *Studia Scientiarum Mathematicarum Hungarica* **19,** 95–113.

[29] Marton, K. (2004). Measure concentration for Euclidean distance in the case of dependent random variables. *Ann. Probab.* **32,** 2526–2544.

[30] Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* **18,** 1269–1283.

[31] Mossel, E. and Roch, S. (2006). Learning nonsingular phylogenies and hidden Markov models. *Ann. Appl. Probab.* **16,** 583–614.

[32] Parkes, D. C., Adams, R. P., Hsu, D. and Zou, J. Y. (2013). Contrastive learning using spectral methods. In *Advances in Neural Information Processing Systems 26*. ed. C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. pp. 2238–2246.

[33] Rio, E. (2000). Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *C. R. Acad. Sci. Paris Sér. I Math.* **330,** 905–908.

[34] Samson, P.-M. (2000). Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.* **28,** 416–461.

[35] Siddiqi, S. M., Boots, B. and Gordon, G. J. (2010). Reduced-rank Hidden Markov Models. In *AISTAT*.

[36] Wagner, R. (2008). Tail estimates for sums of variables sampled by a random walk. *Combinatorics, Probability & Computing* **17,** 307–316.