Uniform Chernoff and Dvoretzky-Kiefer-Wolfowitz-type inequalities for Markov chains and related processes

Aryeh Kontorovich∗†  Roi Weiss∗

May 2, 2014

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.

1 Introduction

1.1 Background
The last decade or so has 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 $\mathcal{L}(X^n_j | X^n_i = x)$ 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 $w, w' \in \mathbb{N}$,

$$\eta_{ij}(y, w, w') = \|\mathcal{L}(X^n_j | X^n_i = yw) - \mathcal{L}(X^n_j | X^n_i = yw')\|_{TV},$$

(1)

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

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

The coefficients $\bar{\eta}_{ij}$, termed $\eta$-mixing coefficients in [21], play a central role in several recent concentration results. Define $\Delta$ to be the upper-triangular $n \times n$ matrix, with $\Delta_{ii} = 1$ and $\Delta_{ij} = 0$ for $i < j$. The matrix $\Delta$ is a geometrically ergodic Markov chain, and hence Proposition 1.2 implies

\[\mathbb{P}(\|X^n - \mathbb{E}[X^n]\|_{TV} > \epsilon) \leq \exp(-n \cdot \eta(n) / \Delta_{max}),\]

for some constant $\eta(n)$ and $\Delta_{max}$.

∗Department of Computer Science, Ben-Gurion University, Beer Sheva, Israel
†This research was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 1141/12).
\( \Delta_{ij} = \bar{\eta}_{ij} \) for 1 \( \leq i < j \leq n \). In 2007, [7] and [21] independently proved that for any \( f : \mathbb{N}^n \rightarrow \mathbb{R} \) with \( \|f\|_{\text{Lip}} \leq 1 \) with respect to the Hamming metric\(^1\), we have

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

where \( \|\Delta\|_p \) is the \( \ell_p \) operator norm ([7] achieve the better constant in the exponent, given here). Earlier, Samson [34] had given a concentration result for convex \( \ell_2 \)-Lipschitz functions \( f : [0,1]^n \rightarrow \mathbb{R} \), 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

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

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

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

for 1-Lipschitz functions \( f \), which Marton [26] had essentially obtained earlier by other means. 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 \( \pi \) and conditional \( s \)th step distribution \( \mathcal{L}(X_s | X_1 = x) \), and define the “inverse mixing time”\(^2\)

\[
\tau_s = \sup_{x \in \mathbb{N}} \|\mathcal{L}(X_s | X_1 = x) - \pi\|_{\text{TV}}.
\]

A simple calculation (Lemma 7) shows that \( \bar{\eta}_{ij} \leq 2\tau_{j-i} \), and thus

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

A rich body of work deals with bounding \( \tau_s \) 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

\[
\tau_s \leq G\theta^{s-1}, \quad s = 1,2,\ldots
\]

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

\(^2\)This terminology is non-standard.
Any finite ergodic Markov chain is geometrically ergodic, and the dependence of $G, \theta$ 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 $\tau_s$ 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 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 “$(G, \theta)$-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.** Let $Y_1, Y_2, \ldots$ be a $\mathbb{N}$-valued hidden Markov chain whose underlying $\mathbb{N}$-valued Markov chain is $(G, \theta)$-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), we have

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

with an identical bound for the other tail.

Although the result in Theorem 1 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 $(X_1, \ldots, X_n)$:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq x), \quad x \in \mathbb{N}.$$
The DKW inequality states that
\[ P \left( \sup_{x \in \mathbb{N}} \left| \hat{F}_n(x) - F(x) \right| > \varepsilon \right) \leq 2 \exp(-2n\varepsilon^2), \quad \varepsilon > 0, n \in \mathbb{N}. \]

We present the following Markovian version of this inequality.

**Theorem 2.** Let \( Y_1, Y_2, \ldots \) be a stationary \( \mathbb{N} \)-valued \((G, \theta)\)-geometrically ergodic Markov or hidden Markov chain with stationary distribution \( \rho \in \mathbb{R}^N \). For \( n \in \mathbb{N} \), define \( \hat{\rho}^{(n)} \in \mathbb{R}^N \) to be the empirical estimate of \( \rho \):
\[ \hat{\rho}^{(n)}_y = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{Y_i = y\}}, \quad y \in \mathbb{N}. \] (6)

Then
\[ P \left( \left\| \rho - \hat{\rho}^{(n)} \right\|_\infty > \frac{\sqrt{1 + 2G\theta}}{n(1-\theta)} + \varepsilon \right) \leq \exp \left( -\frac{n(1-\theta)^2\varepsilon^2}{2G^2} \right), \quad n \in \mathbb{N}, \varepsilon > 0. \] (7)

Note that a naive application of Theorem 1 to each \( \hat{\rho}^{(n)}_y \) individually, combined with the union bound, yields
\[ P \left( \left\| \rho - \hat{\rho}^{(n)} \right\|_\infty > \varepsilon \right) \leq 2\|\rho\|_0 \exp \left( -\frac{n(1-\theta)^2\varepsilon^2}{2G^2} \right), \] where \( \|\rho\|_0 \) is the number of non-zero entries in \( \rho \). The bound in (7) is vacuous for \( \rho \) with infinite support. The assumption that the chain starts in the stationary distribution is not at all restrictive, as shown in Section 2.6.

**Uniform Chernoff bound.** Let \( Y_1, Y_2, \ldots \) be a stationary \( \mathbb{N} \)-valued \((G, \theta)\)-geometrically ergodic Markov or hidden Markov chain as above, and consider the occupation frequency:
\[ \hat{\rho}^{(n)}(E) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{Y_i \in E\}}, \quad E \subseteq \mathbb{N}. \]

A naive application of Theorem 1 might yield a deviation bound along the lines of
\[ P \left( |\rho(E) - \hat{\rho}^{(n)}(E)| > \varepsilon \right) \leq 2|E| \exp \left( -\frac{n(1-\theta)^2\varepsilon^2}{2|E|^2G^2} \right), \] where \( |E| \) is the cardinality of \( E \) and \( \rho \) 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 3.** Define
\[ \Lambda_n(\rho) = \gamma_n(G, \theta) \sum_{\rho_y \geq 1/n} \sqrt{\rho_y} + \min \left\{ \gamma_n(G, \theta) \sum_{\rho_y < 1/n} \sqrt{\rho_y}, \sum_{\rho_y < 1/n} \rho_y \right\}, \quad n \in \mathbb{N}, \]
where
\[ \gamma_n(G, \theta) = \frac{1}{2} \sqrt{1 + 2G\theta} \frac{1}{n(1 - \theta)}. \]

Then:

(a) for all distributions \( \rho \in \mathbb{R}^N \),
\[ \lim_{n \to \infty} \Lambda_n(\rho) = 0, \]

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

We remark that the rate at which \( \Lambda_n(\rho) \) decays to 0 depends on \( \rho \) and may be arbitrarily slow for heavy-tailed distributions. When \( \sum_{y \in N} \sqrt{\rho_y} < \infty \), we get a simpler estimate in (b) via
\[ \Lambda_n(\rho) \leq \gamma_n(G, \theta) \sum_{y \in N} \sqrt{\rho_y}. \]

Again, the stationarity assumption is quite mild (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 independent results on 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 \( A(x \mid y) \). 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 \( \mathbb{N}^n \):
\[ \mathcal{L}(X_1, \ldots, X_n) = p_1(X_1) \prod_{i=1}^{n-1} A(X_{i+1} \mid X_i). \]
**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 \(N \times N\) column-stochastic matrix of emission probabilities. This HMM induces a distribution on \(N^n\) as follows. Let \(X \in N^n\) be distributed according to (8) and define the conditional distribution \(\mathcal{L}(\cdot | X)\) over \(Y \in N^n\):

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

It follows that

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

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 \(\pi \in \mathbb{R}^N\) 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

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

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

\[
\left\| \mathbb{E} \hat{\pi}^{(n)} - \pi \right\|_{TV} \leq \frac{1}{n} \sum_{i=1}^{n} \|\mathcal{L}(X_i) - \pi\|_{TV} \leq \frac{1}{n} \sum_{i=1}^{n} \sum_{x \in N} \mathcal{L}(X_i \mid X_1 = x)p_1(x) - \pi \|_{TV} \leq \frac{1}{n} \sum_{i=1}^{n} \sum_{x \in N} p_1(x) \|\mathcal{L}(X_i \mid X_1 = x) - \pi\|_{TV} \leq \frac{1}{n} \sum_{i=1}^{n} \sum_{x \in N} p_1(x)G^{i-1} = \frac{G}{(1 - \theta)n}.
\]

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

\[
\|E\hat{\rho}(n) - \rho\|_{\text{TV}} \leq \|E\hat{\pi}(n) - \pi\|_{\text{TV}} \leq \frac{G}{(1 - \theta)n}.
\] (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 4.** Let \( X \) and \( X' \) be two Markov chains induced by \((\xi, A)\) and \((\xi', A')\), respectively. For a given emission matrix \( B \), let \( Y \) and \( Y' \) be the hidden Markov chains induced by \((\xi, A, B)\) and \((\xi', A', B)\). Then

\[
\|L_i - L_i'\|_{\text{TV}} \leq \|L_i - L_i'\|_{\text{TV}}, \quad I \subseteq \{1, \ldots, n\}, n \in \mathbb{N}.
\]

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

The proofs of Theorems 2 and 3 will require bounds on \( \|\hat{\rho}(n) - \rho\| \), 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:

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

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

**Lemma 5** (Markov, 1906 [25]). For any two stochastic vectors \( \xi, \psi \in \mathbb{R}^N \), we have

\[
\|A(\xi - \psi)\|_{\text{TV}} \leq \kappa \|\xi - \psi\|_{\text{TV}}.
\]

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

**Corollary 6.** Let \( A \) be a \((G, \theta)\)-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 \( \pi \) be the stationary distribution of \( A \) and \( \xi, \psi \in \mathbb{R}^N \) two point masses. Then

\[
\|A^n\xi - A^n\psi\|_{\text{TV}} \leq \|A^n\xi - \pi\|_{\text{TV}} + \|A^n\psi - \pi\|_{\text{TV}} \leq 2\tau_{n+1} \leq 2G\theta^n.
\]

\[\square\]
2.3 Proof of main inequality

In this section, we prove Theorem 1. The first order of business is to bound the $\eta$-mixing coefficient by the inverse mixing time, and hence in terms of $G$ and $\theta$.

**Lemma 7.** Let $Y$ be a $(G, \theta)$-geometrically ergodic hidden Markov chain and let $\bar{\eta}_{ij}$ and $\tau_n$ be as defined in (1) and (4), respectively. Then

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

**Proof.** Let $X$ be the Markov chain underlying $Y$ and endow $\bar{\eta}_{ij}(X)$ 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 6 yields

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

**Proof of Theorem 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 7, we get

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

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

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

2.4 Proof of the DKW-type inequality

In this section, we prove Theorem 2. Let $Y_1, Y_2, \ldots$ be a stationary $(G, \theta)$-geometrically ergodic hidden Markov chain with stationary distribution $\rho$, and define the $\{0, 1\}$-indicator variables

$$\xi_{i}^{(y)} = \mathbb{1}_{\{Y_i = y\}}, \quad i, y \in \mathbb{N}. \quad (10)$$

Then $\hat{\rho}$, defined in (6), is given by

$$\hat{\rho}_y = \frac{1}{n} \sum_{i=1}^{n} \xi_{i}^{(y)}, \quad i, y \in \mathbb{N}. \quad (11)$$

Then $\hat{\rho}$ is approximated by $\hat{\rho}$ through $\rho$.

Observing that the map $(Y_1, Y_2, \ldots) \mapsto n \|\rho - \hat{\rho}\|_{\infty}$ is 1-Lipschitz under the Hamming metric (Lemma 13), we apply Theorem 1:

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

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

**Lemma 8.**

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

**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]

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

**Proof.** Jensen’s inequality yields

$$(E \|\rho - \hat{\rho}\|_{\infty})^2 \leq E \left[\|\rho - \hat{\rho}\|_{\infty}^2\right].$$

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

$$n^2 Var[\hat{\rho}_y] = E\left(\sum_{y \in \mathbb{N}} (\rho_y - \hat{\rho}_y)^2\right) = \sum_{y \in \mathbb{N}} Var[\hat{\rho}_y]. \quad (12)$$

and

$$E S_{n}^{(y)} = n \rho_y. \quad (13)$$

To bound $E\left(S_{n}^{(y)}\right)^2$, we compute

$$E\left(S_{n}^{(y)}\right)^2 = E\left[\sum_{1 \leq i, j \leq n} \xi_{i}^{(y)} \xi_{j}^{(y)}\right] = \sum_{i=1}^{n} E\left(\xi_{i}^{(y)}\right)^2 + 2 \sum_{1 \leq i < j \leq n} E\left[\xi_{i}^{(y)} \xi_{j}^{(y)}\right] = n \rho_y + 2 \sum_{1 \leq i < j \leq n} E\left[\xi_{i}^{(y)} \xi_{j}^{(y)}\right], \quad (14)$$

9
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
\[
\| \mathcal{L}(Y_i \mid Y_1 = y) - \rho \|_{\infty} \leq G\theta^{i-1}, \quad i, y \in \mathbb{N}.
\]
Indeed, denoting the parameters of \( Y \) by \( (\pi, A, B) \) and letting \( X \) be the underlying Markov chain, we have
\[
\begin{align*}
\| \mathcal{L}(Y_i \mid Y_1 = y_1) - \rho \|_{\infty} &\leq \| \mathcal{L}(Y_i \mid Y_1 = y_1) - \rho \|_{TV} \\
&= \frac{1}{2} \sum_{y_i \in \mathbb{N}} \sum_{x_i \in \mathbb{N}} B_{y_i, x_i} \frac{1}{2} \sum_{y_i \in \mathbb{N} x_i \in \mathbb{N}} B_{y_i, x_i} \left| \mathbb{P}(X_i = x_i \mid Y_1 = y_1) - \pi_{x_i} \right| \\
&\leq \frac{1}{2} \sum_{y_i \in \mathbb{N} x_i \in \mathbb{N}} B_{y_i, x_i} \left| \mathbb{P}(X_i = x_i \mid Y_1 = y_1) - \pi_{x_i} \right| \\
&= \left| \sum_{x_1 \in \mathbb{N}} \mathcal{L}(X_i \mid X_1 = x_1) \mathbb{P}(X_1 = x_1 \mid Y_1 = y_1) - \pi \right|_{TV} \\
&\leq \sup_{x_1 \in \mathbb{N}} \| \mathcal{L}(X_i \mid X_1 = x_1) - \pi \|_{TV} \leq G\theta^{i-1}.
\end{align*}
\]
Hence,
\[
\mathbb{E} \left[ \xi_i^{(y)} \xi_j^{(y)} \right] = \mathbb{P}(Y_i = y, Y_j = y) \\
= \mathbb{P}(Y_1 = y, Y_{j-i+1} = y) \\
= \mathbb{P}(Y_1 = y) \mathbb{P}(Y_{j-i+1} = y \mid Y_1 = y) \\
\leq \rho_y (\rho_y + G\theta^{j-i}),
\]
and therefore
\[
\sum_{1 \leq i < j \leq n} \mathbb{E} \left[ \xi_i^{(y)} \xi_j^{(y)} \right] = \sum_{k=1}^{n-1} (n-k) \mathbb{P}(Y_1 = y) \mathbb{P}(Y_{k+1} = y \mid Y_1 = y) \\
\leq \sum_{k=1}^{n-1} (n-k) \rho_y (\rho_y + G\theta^k) \\
= \frac{n(n-1)}{2} \rho_y^2 + \frac{G\theta}{1-\theta} \left( n - \frac{1-\theta^n}{1-\theta} \right) \rho_y \\
\leq \frac{n(n-1)}{2} \rho_y^2 + n \frac{G\theta}{1-\theta} \rho_y.
\] (15)
Combining (12), (13), (14), and (15), we have
\[
\text{Var}[\hat{\rho}_y] \leq \frac{1}{n^2} \left(n\rho_y + n(n-1)\rho_y^2 + 2n\frac{G\theta}{1-\theta}\rho_y - n^2\rho_y^2\right)
\]
\[
= \frac{\rho_y}{n} \left(1 - \rho_y + \frac{2G\theta}{1-\theta}\right)
\]
\[
\leq \frac{\rho_y}{n(1-\theta)}.
\]
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 \(\ell_2\) norm.

2.5 Proof of the uniform Chernoff bound

In this section, we prove Theorem 3. As before, \(Y_1, Y_2, \ldots\) is a stationary \((G, \theta)\)-geometrically ergodic hidden Markov chain with stationary distribution \(\rho\). Since by Lemma 13 the map \((Y_1, \ldots, Y_n) \mapsto n\|\rho - \hat{\rho}\|_{TV}\) is 1-Lipschitz under the Hamming metric, Theorem 1 applies:

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

As before, the crux of the matter is to bound \(E\|\rho - \hat{\rho}\|_{TV}\). Recall the definition of \(\Lambda_n\) from the statement of Theorem 3.

Lemma 9.

\[
E\|\rho - \hat{\rho}\|_{TV} \leq \Lambda_n.
\]

Remark. This bound is nearly optimal: when the \(Y_i\) are iid, we have \([4, \text{Proposition 3}]\)

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

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

\[
E\|\rho - \hat{\rho}\|_{TV} = \frac{1}{2} \sum_{y: \rho_y < 1/n} E|\rho_y - \hat{\rho}_y| + \frac{1}{2} \sum_{y: \rho_y \geq 1/n} E|\rho_y - \hat{\rho}_y|, \tag{17}
\]

and bounding each term separately. To bound the second term, we note, as in the proof of Lemma 8, that

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

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

\[
nE|\rho_y - \hat{\rho}_y| = E \left| \sum_{i=1}^n \xi_i^{(y)} - n\rho_y \right|
\]
\[
\leq nE|\xi_i^{(y)} - \rho_y|
\]
\[
= 2n\rho_y(1-\rho_y) \leq 2n\rho_y,
\]
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 3. (a) Since obviously

\[
\sum_{\rho_y < 1/n} \rho_y \rightarrow 0, \quad n \to \infty
\]

it suffices to show that

\[
\frac{1}{\sqrt{n}} \sum_{\rho_y \geq 1/n} \sqrt{\rho_y} \rightarrow 0.
\] (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 arbitrary \(\varepsilon > 0\), and let \(N \in \mathbb{N}\) be large enough so that \(\sum_{j \geq N} \rho_j < \varepsilon\). Then

\[
\frac{1}{\sqrt{n}} \sum_{\rho_j \geq 1/n} \sqrt{\rho_j} \leq \frac{1}{\sqrt{n}} \sum_{j \leq N} \sqrt{\rho_j} + \frac{1}{\sqrt{n}} \sum_{j > N, \rho_j \geq 1/n} \sqrt{\rho_j}
\]

\[
\leq \sqrt{\frac{N}{n}} + \frac{1}{\sqrt{n}} \sqrt{\sum_{j \geq 1/n} \rho_j} \leq \sqrt{\frac{N}{n}} + \varepsilon,
\]

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 \(\phi, \psi \in \mathbb{R}^n\),

\[
\|\phi - \psi\|_{TV} = \sup_{E \subseteq \mathbb{N}} |\phi(E) - \psi(E)|.
\]

2.6 The stationarity assumption

For rapidly mixing Markov and hidden Markov chains, the stationarity assumption can easily be relaxed. Indeed, Let \(Y = (Y_1, \ldots, Y_n)\) be a \((G, \theta)\)-geometrically ergodic hidden Markov chain with parameters \((B\pi', A, B)\), where \(\pi' \in \mathbb{R}^N\) is some stochastic vector. If \(Y\) is “nearly stationary,” in the sense that \(\|\pi - \pi\|_{TV}\) is small, a simple dimension-free bound on the statistical distance between \(Y\) and its stationary version is available.

Theorem 10. Let \(Y' = (Y'_1, \ldots, Y'_n)\) be the stationary version of \(Y\) — i.e., an HMM with parameters \((B\pi, A, B)\), where \(\pi\) is the stationary distribution of the kernel \(A\). Then

\[
\|\mathcal{L}(Y) - \mathcal{L}(Y')\|_{TV} \leq \|\pi - \pi'\|_{TV}.
\]

\(^3\)This elegant proof is due to Asaf Shachar. Andrew Barron points out that (19) may be easily derived from Lebesgue’s dominated convergence theorem.
First, we prove an analogous result for Markov chains.

**Lemma 11.** Let $A$ be Markov kernel and $\xi, \xi' \in \mathbb{R}^N$ two arbitrary stochastic vectors. Let $X = (X_1, \ldots, X_n)$ and $X' = (X'_1, \ldots, X'_n)$ be the Markov chains induced by $(\xi, A)$ and $(\xi', A)$, respectively. Then

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

**Proof.**

$$\|L(X) - L(X')\|_{TV} = \frac{1}{2} \sum_{x \in \mathbb{N}^n} |(\xi_{x_1} - \xi'_{x_1}) A_{x_2, x_1} \cdots A_{x_n, x_{n-1}}|$$

$$= \frac{1}{2} \sum_{x \in \mathbb{N}^n} A_{x_2, x_1} \cdots A_{x_n, x_{n-1}} |\xi_{x_1} - \xi'_{x_1}|$$

$$= \frac{1}{2} \sum_{x_1 \in \mathbb{N}} |\xi_{x_1} - \xi'_{x_1}| = \|\xi - \xi'\|_{TV}.$$ 

**Proof of Theorem 10.** Lemma 4 lets us restrict our attention to the underlying Markov chains $X$ and $X'$, respectively:

$$\|L(Y_{1 \leq i \leq n}) - L(Y'_{1 \leq i \leq n})\|_{TV} \leq \|L(X_{1 \leq i \leq n}) - L(X'_{1 \leq i \leq n})\|_{TV}$$

$$= \|L(X_1) - L(X'_1)\|_{TV} = \|\pi - \pi'\|_{TV},$$

where the first identity follows from Lemma 11.

**Corollary 12.** Let $Y_1, Y_2, \ldots$ be a (not necessarily stationary) $\mathbb{N}$-valued $(G, \theta)$-geometrically ergodic hidden Markov chain with stationary distribution $\rho = B\pi$ and initial distribution $\rho' = B\pi$. Then the deviation bounds stated in Theorems 2 and 3 hold with an additive correction of $\|\pi - \pi'\|_{TV}$ on the right-hand side.

Letting a $(G, \theta)$-geometrically ergodic chain run for $s$ steps before starting the estimation ensures that $\|\pi - \pi'\|_{TV} \leq G\theta^s$.

### 2.7 Auxiliary lemma

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

**Lemma 13.** Suppose $n \in \mathbb{N}$ and $p \in \mathbb{R}^N$ is a distribution. Define the functions $g, h : \mathbb{N}^n \to \mathbb{R}$ by

$$g(x) = \sup_{j \in \mathbb{N}} \left| np_j - \sum_{i=1}^n \mathbb{1}_{\{x_i = j\}} \right|, \quad x \in \mathbb{N}^n,$$

$$h(x) = \sum_{j \in \mathbb{N}} \left| np_j - \sum_{i=1}^n \mathbb{1}_{\{x_i = j\}} \right|, \quad x \in \mathbb{N}^n.$$
Then $\|g\|_{\text{Lip}} \leq 1$ and $\|h\|_{\text{Lip}} \leq 2$ with respect to the Hamming metric:

\[
\begin{align*}
|g(x) - g(y)| & \leq d(x, y), \\
|h(x) - h(y)| & \leq 2d(x, y)
\end{align*}
\]

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

**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, $\hat{n}_j(x) = \sum_{i=1}^n 1_{\{x_i = j\}}$. Now suppose $x, y \in \mathbb{N}^n$ differ only in coordinate $k$, with $x_k = a$ and $y_k = b$. Then

\[
\begin{align*}
\hat{n}_j(x) - \hat{n}_j(y) & = \sum_{j \in \mathbb{N}} |np_j - \hat{n}_j(x)| - \sum_{j \in \mathbb{N}} |np_j - \hat{n}_j(y)| \\
& = (|np_a - \hat{n}_a(x)| + |np_b - \hat{n}_b(x)|) - (|np_a - \hat{n}_a(y)| + |np_b - \hat{n}_b(y)|) \\
& = (|np_a - \hat{n}_a(x)| + |np_b - \hat{n}_b(x)|) - (|np_a - (\hat{n}_a(x) - 1)| + |np_b - (\hat{n}_b(x) + 1)|) \\
& \leq |np_a - \hat{n}_a(x)| - |np_a - (\hat{n}_a(x) - 1)| + |np_b - \hat{n}_b(x)| - |np_b - (\hat{n}_b(x) + 1)| \\
& \leq 2.
\end{align*}
\]

\[\square\]

**Acknowledgments**

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

**References**

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

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

[3] Animashree Anandkumar, Daniel Hsu, and Sham M. Kakade. A method of moments for mixture models and hidden Markov models. In *COLT*, 2012.

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

[5] S. G. Bobkov and F. Götze. Concentration of empirical distribution functions with applications to non-i.i.d. models. *Bernoulli*, 16(4):1385–1414, 2010.

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

[7] Jean-René Chazottes, Pierre Collet, Christof Külske, and Frank Redig. Concentration inequalities for random fields via coupling. *Probability Theory and Related Fields*, 137(1-2):201–225, 2007.
[8] Jean-René Chazottes and Frank Redig. Concentration inequalities for Markov processes via coupling. *Electron. J. Probab.*, 14:no. 40, 1162–1180, 2009.

[9] Kai-Min Chung, Henry Lam, Zhenming Liu, and Michael Mitzenmacher. Chernoff-hoeffding bounds for Markov chains: Generalized and simplified. In *STACS*, pages 124–135, 2012.

[10] Persi Diaconis and Laurent Saloff-Coste. Logarithmic Sobolev inequalities for finite Markov chains. *Ann. Appl. Probab.*, 6(3):695–750, 1996.

[11] Persi Diaconis and Laurent Saloff-Coste. Nash inequalities for finite Markov chains. *Journal of Theoretical Probability*, 9:459–510, 1996.

[12] Ian H. Dinwoodie. A probability inequality for the occupation measure of a reversible Markov chain. *Ann. Appl. Probab.*, 5(1):37–43, 1995.

[13] Ian H. Dinwoodie. Expectations for nonreversible Markov chains. *Journal of Mathematical Analysis and Applications*, 220(2):585 – 596, 1998.

[14] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27:642–669, 1956.

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

[16] David Gillman. A Chernoff bound for random walks on expander graphs. *SIAM J. Comput.*, 27(4):1203–1220, 1998.

[17] Daniel Hsu, Sham M. Kakade, and Tong Zhang. A spectral algorithm for learning hidden Markov models. In *COLT*, 2009.

[18] Nabil Kahale. Large deviation bounds for Markov chains. *Combinatorics, Probability & Computing*, 6(4):465–474, 1997.

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

[20] Aryeh Kontorovich, Boaz Nadler, and Roi Weiss. On learning parametric-output hmms. In *ICML (3)*, pages 702–710, 2013.

[21] Leonid (Aryeh) Kontorovich and Kavita Ramanan. Concentration Inequalities for Dependent Random Variables via the Martingale Method. *Ann. Probab.*, 36(6):2126–2158, 2008.

[22] I. Kontoyiannis and S.P. Meyn. Geometric ergodicity and the spectral gap of non-reversible Markov chains. *Probability Theory and Related Fields*, 154(1-2):327–339, 2012.

[23] Carlos A. León and François Perron. Optimal Hoeffding bounds for discrete reversible Markov chains. *Ann. Appl. Probab.*, 14(2):958–970, 2004.
[24] Pascal Lezaud. Chernoff-type bound for finite Markov chains. *Ann. Appl. Probab.*, 8(3):849–867, 1998.

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

[26] Katalin Marton. Bounding $\tilde{d}$-distance by informational divergence: a method to prove measure concentration. *Ann. Probab.*, 24(2):857–866, 1996.

[27] Katalin Marton. Measure concentration for a class of random processes. *Probability Theory and Related Fields*, 110(3):427–439, 1998.

[28] Katalin Marton. Measure concentration and strong mixing. *Studia Scientiarum Mathematicarum Hungarica*, 19(1-2):95–113, 2003.

[29] Katalin Marton. Measure concentration for Euclidean distance in the case of dependent random variables. *Ann. Probab.*, 32(3):2526–2544, 2004.

[30] Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283, 1990.

[31] Elchanan Mossel and Sébastien Roch. Learning nonsingular phylogenies and hidden Markov models. *Ann. Appl. Probab.*, 16(2):583–614, 2006.

[32] David C. Parkes, Ryan P. Adams, Daniel Hsu, and James Y. Zou. Contrastive learning using spectral methods. In Christopher J. C. Burges, Léon Bottou, Max Welling, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2238–2246. 2013.

[33] Emmanuel Rio. Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(10):905–908, 2000.

[34] Paul-Marie Samson. Concentration of measure inequalities for Markov chains and $\Phi$-mixing processes. *Ann. Probab.*, 28(1):416–461, 2000.

[35] Sajid M. Siddiqi, Byron Boots, and Geoffrey J. Gordon. Reduced-rank Hidden Markov Models. In *AISTAT*, 2010.

[36] Roy Wagner. Tail estimates for sums of variables sampled by a random walk. *Combinatorics, Probability & Computing*, 17(2):307–316, 2008.