Markov Processes Relat. Fields **18**, 613–638 (2012)

Markov
Processes
and
Related Fields

Obtaining Measure Concentration from Markov Contraction

A. Kontorovich

Department of Computer Science, Ben-Gurion University, Beer Sheva, Israel

Received September 12, 2011, revised February 20, 2012

Abstract. Concentration bounds for non-product, non-Haar measures are fairly recent: the first such result was obtained for contracting Markov chains by Marton in 1996 via the coupling method. The work that followed, with few exceptions, also used coupling. Although this technique is of unquestionable utility as a theoretical tool, it is not always simple to apply. As an alternative to coupling, we use the elementary Markov contraction lemma to obtain simple, useful, and apparently novel concentration results for various Markov-type processes. Our technique consists of expressing probabilities as matrix products and applying Markov contraction to these expressions; thus it is fairly general and holds the potential to yield further results in this vein.

Keywords: Markov-type processes, contraction, concentration inequality, η-mixing coeficients

AMS Subject Classification: 60J05, 60J10, 60F10

1. Introduction

1.1. Background

In 1996 Marton [25] published a concentration inequality for contracting Markov chains — apparently, the first such result for a non-product, non-Haar measure. In the decade that followed, Marton and others continued to deepen and broaden a key insight: analogues of the Azuma – Hoeffding – McDiarmid inequality [2, 14, 30] for independent random variables may be obtained for dependent ones, provided a strong mixing condition holds.

To recall, the aforementioned inequality implies that if μ is a product distribution on Ω^n and $f : \Omega^n \to \mathbb{R}$ satisfies $\|f\|_{\mathrm{Lip}} \le 1$ under the Hamming metric, we have

$$\mu\{|f - \mu f| > t\} \le 2\exp(-2t^2/n). \tag{1.1}$$

In [25], Marton pioneered the transportation method for proving concentration inequalities. This technique is in principle applicable to arbitrary nonproduct measures, and when applied to Markov chains μ with contraction coefficient $\theta < 1$, it yields

$$\mu\{|f - M_f| > t\} \le 2\exp\Big[-\frac{2}{n}\Big(t(1-\theta) - \sqrt{\frac{\log 2}{2n}}\Big)\Big]^2, \tag{1.2}$$

where M_f is a μ-median of f. Since product distributions are degenerate cases of Markov chains (with $\theta = 0$), Marton's result is a powerful generalization of (1.1).

The Markov contractivity condition $\theta < 1$ implies strong mixing, and in a series of papers [26–28], Marton gave other concentration results for dependent variables under various metrics and types of mixing. In particular, Theorem 2 of [26] gives a generic mixing condition which implies a transportation inequality and therefore concentration.

Further progress in obtaining concentration from mixing was made, among others, in [6–8, 20, 21, 34, 35]. Using Stein's method for exchangeable pairs, Chatterjee [6] obtained an elegant concentration inequality in terms of a Dobrushin – Shlosman type contractivity condition. Samson [35] was apparently the first to use explicit mixing coefficients in a concentration result. Since these are central to this paper we define them without further delay; the (standard) notation is clarified in Section 1.3.

Let μ be the joint distribution of $(X_1, \ldots, X_n)$, $X_i \in \Omega$. For $1 \le i < j \le n$ and $x \in \Omega^i$, we denote by

$$\mu\big((X_j, \ldots, X_n) \mid (X_1, \ldots, X_i) = x\big)$$

the distribution of $(X_j, \ldots, X_n)$ conditioned on $(X_1, \ldots, X_i) = x$. For $y \in \Omega^{i-1}$ and $w, w' \in \Omega$, define

$$\begin{aligned}\eta_{ij}(y, w, w') = \big\|&\mu\big((X_j, \ldots, X_n) \mid (X_1, \ldots, X_i) = yw\big) \\ &- \mu\big((X_j, \ldots, X_n) \mid (X_1, \ldots, X_i) = yw'\big)\big\|_{\mathrm{TV}},\end{aligned}$$

and

$$\bar{\eta}_{ij} = \sup_{y \in \Omega^{i-1}, w, w' \in \Omega} \eta_{ij}(y, w, w'). \tag{1.3}$$

The coefficients $\bar{\eta}_{ij}$, termed *η-mixing coefficients*[1] in [20], play a key role in several recent concentration results. Define Γ and Δ to be upper-triangular

[1] That choice of terminology is perhaps suboptimal in light of the unrelated notion of *η-weak dependence* of Doukhan et al. [11], but the sufficiently distinct contexts should prevent confusion.

$n \times n$ matrices, with $\Gamma_{ii} = \Delta_{ii} = 1$ and

$$\Gamma_{ij} = \sqrt{\bar{\eta}_{ij}}, \qquad \Delta_{ij} = \bar{\eta}_{ij}$$

for $1 \le i < j \le n$.

In 2000, Samson [35] proved that any distribution μ on $[0,1]^n$ and any convex $f : [0,1]^n \to \mathbb{R}$ with $\|f\|_{\mathrm{Lip}} \le 1$ (with respect to ℓ_2) satisfy

$$\mu\{|f - \mu f| > t\} \le 2 \exp\Big(-\frac{t^2}{2\|\Gamma\|_2^2}\Big) \tag{1.4}$$

where $\|\Gamma\|_2$ is the ℓ_2 operator norm.

In 2007, almost synchronously and using different techniques, Chazottes et al. [7] and the author with K. Ramanan [20] showed that any distribution μ on Ω^n and any $f : \Omega^n \to \mathbb{R}$ with $\sqrt{n}\,\|f\|_{\mathrm{Lip}} \le 1$ (with respect to the Hamming metric) satisfy

$$\mu\{|f - \mu f| > t\} \le 2 \exp\Big(-\frac{t^2}{2\|\Delta\|_\infty^2}\Big) \tag{1.5}$$

where $\|\Delta\|_\infty$ is the ℓ_∞ operator norm ($\|\Delta\|_\infty$ may be replaced by $\|\Delta\|_2$ and in [7] a better constant in the exponent is achieved). More recently, Chazottes and Redig [8] obtained polynomial concentration bounds in terms of moments of the coupling time. The contraction condition may be cast more generally as a curvature [31] or a "metric ergodicity" [16] condition. Some recent results on transportation methods in concentration include [3, 9, 13]; a survey may be found in [17].

The results (1.4) and (1.5) are not readily comparable as they hold in different spaces for different metrics with different normalization, and the former requires convexity. They share the feature of establishing concentration for a wide class of measures, in terms of the natural mixing coefficients $\bar{\eta}_{ij}$. Indeed, since

$$\|\Delta\|_\infty = \max_{1 \le i < n} (1 + \bar{\eta}_{i,i} + \bar{\eta}_{i,i+1} + \ldots + \bar{\eta}_{i,n}) \tag{1.6}$$

and by the Geršgorin disc theorem [15]

$$\|\Gamma\|_2^2 = \lambda_{\max}(\Gamma^{\mathsf{T}}\Gamma) \le \max_{1 \le i \le n} \sum_{j=1}^{n} (\Gamma^{\mathsf{T}}\Gamma)_{ij},$$

suitable upper estimates on $\bar{\eta}_{ij}$ provide bounds for $\|\Gamma\|_2$ and $\|\Delta\|_\infty$.

Aside from the straightforward observation (due to Samson) that the η-mixing coefficients are bounded by the ϕ-mixing ones (see [5]), we are only aware of a few of cases where simple, readily computed estimates on $\bar{\eta}_{ij}$ are given. In particular, Samson [35] controls $\bar{\eta}_{ij}$ by the contraction coefficients of a Markov chain, and Chazottes et al. [7] give some estimates on $\bar{\eta}_{ij}$ for various temperature regimes of Gibbs random fields. The estimates quoted above are obtained via the coupling method — which, while powerful, often requires some ingenuity to construct the requisite joint distribution, even in the simple case of a Markov chain [25, 35]. In some cases, the coupling may even elude explicit construction [7].

As the random processes of interest become more complex, it becomes progressively more difficult to obtain estimates on mixing coefficients via coupling. We are particularly interested in examining the η-mixing of several Markov-type processes, motivated by statistical and computer science applications. Hidden Markov Models (HMMs) have been used in natural language processing [23, 33] and signal processing [29] for decades, with considerable success. Concentration bounds for Markov Chains (and more generally, HMMs) have implications in machine learning and empirical process theory [12, 18]. A Markov-type process called the *Markov marginal process* (MMP) in [19] underlies adaptive Markov Chain Monte Carlo simulations [1]; these evolve according to an inhomogeneous Markov kernel, which in addition to time also depends on the path history. In a forthcoming work, A. Brockwell and the author give strong laws of large numbers for MMPs in terms of the η-mixing coefficients. Random processes indexed by trees have been attracting the attention of probability theorists for some time [4, 32], and the principal technical contribution of this paper is a bound on $\bar{\eta}_{ij}$ for these types of processes.

Remark 1.1. Although our results rely on Markov contraction (Lemma 2.1) rather than on the coupling method, on some level, the distinction is semantic. The novelty of our method lies in (i) avoiding any constructions (implicit or explicit) of joint distributions (ii) rewriting complicated sums as simple(r) matrix and tensor products (iii) applying Lemma 2.1 to the latter expressions. Thus it seems that our method is sufficiently different from classical coupling techniques to merit the terminological distinction.

1.2. Main results

In this paper we present estimates on the η-mixing coefficients $\bar{\eta}_{ij}$ defined in (1.3), for the various Markov-type processes mentioned above. These bounds immediately imply concentration inequalities for a wide class of metrics and measures, via (1.4) and (1.5).

The precise statements of the results require preliminary definitions and are postponed until later sections. The main technical contribution of this paper is Theorem 4.1, which bounds η-mixing coefficients for Markov-tree processes, yielding what appears to be the first concentration of measure result for these. However, we give equal priority to the goal of presenting Markov contraction as a versatile new method for bounding $\bar{\eta}_{ij}$. The nature of the bounds is to control $\bar{\eta}_{ij}$ — a global function of the distribution μ — by some local, easily computed contraction coefficients of μ. For example, let μ be an inhomogeneous Markov chain defined by the transition kernels $\{p_i : 0 \le i < n\}$, which induces a distribution on Ω^n by

$$\mu(x) = p_0(x_1) \prod_{i=1}^{n-1} p_i(x_{i+1} \mid x_i), \quad x \in \Omega^n.$$

Define the ith *contraction coefficient*:

$$\theta_i = \sup_{y,y' \in \Omega} \|p_i(\cdot \mid y) - p_i(\cdot \mid y')\|_{\mathrm{TV}}, \quad 1 \le i < n. \tag{1.7}$$

This quantity turns out to control the η-mixing coefficients for μ:

$$\bar{\eta}_{ij} \le \theta_i \theta_{i+1} \ldots \theta_{j-1},$$

a fact which is proved in [35] using coupling. In [20] we gave an (arguably simpler) alternative proof, which paves the way for the several new results presented here.

This paper is organized as follows. In Section 1.3 we summarise some basic notation used throughout the paper. Some auxiliary lemmas are given in Section 2. The remaining three sections deal with bounding $\bar{\eta}_{ij}$ for Markov chains (directed and undirected) and Markov tree processes.

1.3. Notation and definitions

Since the contribution of this paper is not measure-theoretic in nature, we henceforth take Ω to be a finite set. Extensions to the countable case are quite straightforward [20] and the continuous case, under mild assumptions, is not much more difficult [18, 19].

We use the terms *measure* and *distribution* interchangeably; all measures are probabilities unless noted otherwise. If μ is a measure on Ω^n and $f : \Omega^n \to \mathbb{R}$, we use the standard notation

$$\mu f = \int_{\Omega^n} f \, d\mu$$

and write

$$\mu\{|f - \mu f| > t\}$$

as a shorthand for

$$\mu\big(\{x \in \Omega^n : |f(x) - \mu f| > t\}\big).$$

The (unnormalized) Hamming metric on Ω^n is defined by

$$d(x,y) = \sum_{i=1}^{n} \mathbb{1}\{x_i \neq y_i\}, \quad x,y \in \Omega^n, \tag{1.8}$$

where the indicator variable $\mathbb{1}\{\cdot\}$ assigns 0-1 truth values to the predicate in $\{\cdot\}$.

The Lipschitz constant of a function, with respect to some metric d, is defined by

$$\|f\|_{\mathrm{Lip}} = \sup_{x \neq y \in \Omega^n} \frac{|f(x) - f(y)|}{d(x,y)}.$$

Random variables are capitalized (X), specified sequences are written in lowercase ($x \in \Omega^n$), the shorthand $X_i^j = (X_i, \dots, X_j)$ is used for all sequences, and sequence concatenation is denoted multiplicatively: $x_i^j x_{j+1}^k = x_i^k$. Sums will range over the entire space of the summation variable; thus $\sum_{x_i^j} f(x_i^j)$ stands for

$$\sum_{x_i^j \in \Omega^{j-i+1}} f(x_i^j).$$

By convention, when $i > j$, we define

$$\sum_{x_i^j} f(x_i^j) \equiv f(\varepsilon)$$

where ε is the null sequence. Products of spaces and measures are denoted by $\otimes$.

The *total variation* norm of a signed measure ν on Ω^n (i.e., vector $\nu \in \mathbb{R}^{\Omega^n}$) is defined by

$$\|\nu\|_{\mathrm{TV}} = \frac{1}{2}\|\nu\|_1 = \frac{1}{2} \sum_{x \in \Omega^n} |\nu(x)|$$

(the factor of $1/2$ is not entirely standard). For readability, we will drop the subscript TV from the norm; thus everywhere in the sequel, $\|\cdot\|$ will mean $\|\cdot\|_{\mathrm{TV}}$.

A signed measure ν on a set $\mathcal{X}$ is called *balanced* if $\nu(\mathcal{X}) = 0$. Departing from standard convention, our stochastic matrices will be column- (as opposed to row-) stochastic. We will use $|\cdot|$ to denote set cardinalities, and write $[n]$ for the set $\{1, \dots, n\}$.

2. Contraction and tensorization

Our method for bounding η-mixing coefficients rests on the following simple result:

Lemma 2.1. *Let $P : \mathbb{R}^\Omega \to \mathbb{R}^\Omega$ be a Markov operator:*

$$(P\nu)(x) = \sum_{y\in\Omega} P(x \mid y)\,\nu(y),$$

where $P(x \mid y) \geq 0$ and $\sum_{x\in\Omega} P(x \mid y) = 1$. Define the contraction coefficient of P as above:

$$\theta = \max_{y,y'\in\Omega} \|P(\cdot \mid y) - P(\cdot \mid y')\|.$$

Then

$$\|P\nu\| \leq \theta\|\nu\|$$

for any balanced signed measure ν on Ω (i.e., $\nu \in \mathbb{R}^\Omega$ with $\sum_{x\in\Omega} \nu(x) = 0$).

This result is sometimes credited to Dobrushin [10]; the quantity θ has been referred to in the literature, alternatively, as the *Doeblin contraction* or *Dobrushin ergodicity* coefficient. However, the observation apparently goes as far back as Markov himself [24] (see [20] for a proof), so it seems appropriate to refer to the result above as the *Markov contraction lemma.*

Another important property of the total variation norm is that it tensorizes, in the following way:

Lemma 2.2. *Consider two finite sets $\mathcal{X}, \mathcal{Y}$, with probability measures p, p' on $\mathcal{X}$ and q, q' on $\mathcal{Y}$. Then*

$$\|p \otimes q - p' \otimes q'\| \leq \|p - p'\| + \|q - q'\| - \|p - p'\|\|q - q'\|.$$

This fact seems to be folklore knowledge; we were not able to locate it in published literature. A non-coupling proof is given in [18], but we will give the simpler coupling proof here.

Proof of Lemma 2.2. Recall that

$$\|r - r'\| = \inf \mathbf{P}\{Z \neq Z'\}$$

where the infimum is over all *couplings* of r and r' — i.e., joint distributions on (Z, Z') with respective marginals r and r'. A coupling achieving the infimum is called *optimal.* Let X, X', Y, Y' be random variables with distributions p, p', q, q', respectively. Define π be an optimal coupling of p and p', and define similarly π' for p', q'. Notice that $\pi \otimes \pi'$ is a (not necessarily optimal) coupling of $p \otimes q$ and $p' \otimes q'$. Then

$$\begin{aligned}
\|p \otimes q - q' \otimes p'\| &= \inf \mathbf{P}\{(X, Y) \neq (X', Y')\} \\
&= \inf \big[\mathbf{P}\{X \neq X'\} + \mathbf{P}\{Y \neq Y'\} - \mathbf{P}\{X \neq X', Y \neq Y'\}\big] \\
&\leq \pi\{X \neq X'\} + \pi'\{Y \neq Y'\} - \pi\{X \neq X'\}\pi'\{Y \neq Y'\} \\
&= \|p - p'\| + \|q - q'\| - \|p - p'\|\|q - q'\|.
\end{aligned}$$

□

3. Markov chains

3.1. Directed

Technically, this section might be considered superfluous, since this result has already appeared in [20], and is strictly generalized in later sections. However, we find it instructive to work out the simple Markov case as it provides the cleanest illustration of our technique.

Let μ be an inhomogeneous Markov measure on Ω^n, induced by the kernels p_0 and $p_i(\cdot \mid \cdot)$, $1 \le i < n$. Thus,

$$\mu(x) = p_0(x_1) \prod_{i=1}^{n-1} p_i(x_{i+1} \mid x_i).$$

The ith contraction coefficient θ_i is defined as in (1.7). As stated in the Introduction, Markov contraction provides an estimate on estimate η-mixing:

Theorem 3.1.

$$\bar{\eta}_{ij} \le \theta_i \theta_{i+1} \dots \theta_{j-1}.$$

Proof. Fix $1 \le i < j \le n$ and $y_1^{i-1} \in \Omega^{i-1}$, $w_i, w_i' \in \Omega$. Then

$$\eta_{ij}(y, w, w') = \frac{1}{2} \sum_{x_j^n} \left| \mu(x_j^n \mid y_1^{i-1} w_i) - \mu(x_j^n \mid y_1^{i-1} w_i') \right| = \frac{1}{2} \sum_{x_j^n} \pi(x_j^n) |\zeta(x_j)|$$

where

$$\pi(u_k^l) = \prod_{t=k}^{l-1} p_t(u_{t+1} \mid u_t)$$

and

$$\zeta(x_j) = \begin{cases} \sum_{z_{i+1}^{j-1}} p_{j-1}(x_j \mid z_{j-1}) \pi(z_{i+1}^{j-1}) \big(p_i(z_{i+1} \mid w_i) - p_i(z_{i+1} \mid w_i') \big), & j - i > 1, \\ p_i(x_j \mid w_i) - p_i(x_j \mid w_i'), & j - i = 1. \end{cases} \tag{3.1}$$

Define $\mathbf{h} \in \mathbb{R}^\Omega$ by $\mathbf{h}_v = p_i(v \mid w_i) - p_i(v \mid w_i')$ and $P^{(k)} \in \mathbb{R}^{\Omega \times \Omega}$ by $P^{(k)}_{u,v} = p_k(u \mid v)$. Likewise, define $\mathbf{z} \in \mathbb{R}^\Omega$ by $\mathbf{z}_v = \zeta(v)$. It follows that

$$\mathbf{z} = P^{(j-1)} P^{(j-2)} \cdots P^{(i+2)} P^{(i+1)} \mathbf{h}.$$

Therefore,

$$\eta_{ij}(y, w, w') = \frac{1}{2} \sum_{x_j^n} \pi(x_j^n) |\mathbf{z}_{x_j}| = \frac{1}{2} \sum_{x_j} |\mathbf{z}_{x_j}| \sum_{x_{j+1}^n} \pi(x_j^n) = \frac{1}{2} \sum_{x_j} |\mathbf{z}_{x_j}| = \|\mathbf{z}\|.$$

The claim follows by (repeated applications of) the Markov contraction lemma. □

The reader may wish to compare this proof with Samson's [35].

Remark 3.1. Aside from simplicity, another advantage of the Markov contraction method is its precision. Namely, in the proof above, note that equality is maintained until the very end, where the Markov contraction lemma is invoked to bound $\|\mathbf{z}\|$. This means that more delicate (for example, spectral) estimates on $\|P^{(j-1)}P^{(j-2)}\dots P^{(i+2)}P^{(i+1)}\mathbf{h}\|$ translate directly into tighter concentration bounds.

3.2. Undirected

In this section we analyze Markov chains under a different parametrization, in an "undirected graphical model" setting [22]. For any graph $G = (V, E)$, where $|V| = n$ and the maximal cliques have size 2 (i.e., are edges), we can define a measure on $\Omega^V = \Omega^n$ as follows

$$\mu(x) = \frac{\prod_{(i,j)\in E}\psi_{ij}(x_i, x_j)}{\sum_{x'\in\Omega^n}\prod_{(i,j)\in E}\psi_{ij}(x'_i, x'_j)}, \quad x \in \Omega^n \tag{3.2}$$

for some nonnegative "potential functions" ψ_{ij}.

Consider the very simple case of chain graphs; any such measure is a Markov measure on Ω^n. We can relate the induced Markov transition kernel $p_i(\cdot \mid \cdot)$ to the random field measure μ as follows:

$$p_i(x \mid y) = \frac{\sum_{v_1^{i-1}}\sum_{z_{i+2}^n}\mu(vyxz)}{\sum_{x'\in\Omega}\sum_{v'^{i-1}_1}\sum_{z'^n_{i+2}}\mu(v'yx'z')}, \quad x, y \in \Omega.$$

Our goal is to bound the ith contraction coefficient θ_i of the Markov chain in terms of ψ_{ij}. We claim a simple relationship between θ_i and ψ_{ij}:

Theorem 3.2.

$$\theta_i \le \frac{R_i - r_i}{R_i + r_i}, \quad 1 \le i < n \tag{3.3}$$

where

$$R_i = \max_{x,y\in\Omega}\psi_{i,i+1}(x, y)$$

and

$$r_i = \min_{x,y\in\Omega}\psi_{i,i+1}(x, y).$$

We will need the following lemma, which may be of independent interest:

Lemma 3.1. *For $n \in \mathbb{N}$ and $0 \le r \le R$, consider the vectors $\alpha \in [0,\infty)^n$ and $f, g \in [r,R]^n$. Then*

$$\frac{1}{2}\sum_{i=1}^{n}\left|\frac{\alpha_i f_i}{\sum_{j=1}^{n}\alpha_j f_j} - \frac{\alpha_i g_i}{\sum_{j=1}^{n}\alpha_j g_j}\right| \le \frac{R-r}{R+r}.$$

Remark. This lemma was proved together with Roi Weiss.

Proof. Assume for now $r > 0$ (the general case will follow by continuity). Since

$$\sum_{i=1}^{n}|x_i - y_i| = \max_{\beta\in\{-1,1\}^n}\sum_{i=1}^{n}\beta_i(x_i - y_i)$$

proving our claim is equivalent to showing that

$$\sum_i \beta_i\Big(\frac{\alpha_i f_i}{\sum_j \alpha_j f_j} - \frac{\alpha_i g_i}{\sum_j a_j g_j}\Big) \le 2\frac{R-r}{R+r} \tag{3.4}$$

holds for all $f, g \in [r,R]^n$, $\alpha \in [0,\infty)^n$, and $\beta \in \{-1,1\}^n$. For fixed α, β, define the function $F_{\alpha,\beta} : [r,R]^{2n} \to \mathbb{R}$ by putting $F_{\alpha,\beta}(f,g)$ equal to the left-hand side of (3.4). If $F_{\alpha,\beta}$ achieves an extremum somewhere on $[r,R]^{2n}$, its gradient must vanish there. Let us compute this gradient:

$$\frac{\partial F_{\alpha,\beta}}{\partial f_i} = \frac{\alpha_i\sum_{j\ne i}\alpha_j(\beta_i-\beta_j)f_j}{(\sum_{k=1}^{n}\alpha_i f_i)^2},$$
$$\frac{\partial F_{\alpha,\beta}}{\partial g_i} = \frac{\alpha_i\sum_{j\ne i}\alpha_j(\beta_i-\beta_j)g_j}{(\sum_{k=1}^{n}\alpha_i g_i)^2}.$$

Solving for $\nabla F_{\alpha,\beta} \equiv 0$, we get that the latter holds whenever

$$f_a = \frac{\sum_{j\notin\{a,b\}}\alpha_j(\beta_b-\beta_j)f_j}{\alpha_a(\beta_a-\beta_b)}, \qquad f_b = \frac{\sum_{j\notin\{a,b\}}\alpha_j(\beta_j-\beta_a)f_j}{\alpha_b(\beta_a-\beta_b)},$$
$$g_c = \frac{\sum_{j\notin\{c,d\}}\alpha_j(\beta_d-\beta_j)g_j}{\alpha_c(\beta_c-\beta_d)}, \qquad g_d = \frac{\sum_{j\notin\{c,d\}}\alpha_j(\beta_j-\beta_c)g_j}{\alpha_d(\beta_c-\beta_d)}$$

for some $a \ne b$ and $c \ne d$. Since the expressions above are undefined for $\beta_a = \beta_b$ or $\beta_c = \beta_d$, we may assume that neither of these holds. We claim that $f_a \le 0$ for $a \ne b$ and $\beta_a \ne \beta_b$; this is easily verified by substituting $\beta_a = 1$, $\beta_b = -1$ and $\beta_a = -1$, $\beta_b = 1$. (A similar observation holds for g_c.) We conclude that $\nabla F_{\alpha,\beta}$ does not vanish anywhere on $[r,R]^{2n}$, and therefore the function must achieve its extreme values on the boundary of this region.

In light of the above, it suffices to consider $f, g \in \{r,R\}^n$. Consider the expression

$$\mathcal{L} = \sum_i \beta_i\Big(\frac{\alpha_i f_i}{A} - \frac{\alpha_i g_i}{B}\Big), \tag{3.5}$$

where $f, g \in \{r, R\}^n$, $\alpha \in [0, \infty)^n$, $\beta \in \{-1, 1\}^n$, and $A = \sum_j \alpha_j f_j$, $B = \sum_j \alpha_j g_j$. Keeping A and B constant, we seek an α that maximizes $\mathcal{L}$. The latter is a linearly constrained linear program, and thus its maximal value(s) are attained at the extreme points of the feasible region $E = E(A, B, f, g) \subset \mathbb{R}^n$, given by

$$E = \big\{\alpha \in [0, \infty)^n : \langle \alpha, f \rangle = A, \langle \alpha, g \rangle = B\big\}.$$

Recalling that $z \in E$ is an extreme point if and only if z cannot be expressed as $z = \lambda x + (1 - \lambda) y$ for $x, y \in E$ and $0 < \lambda < 1$, it is straightforward to verify that the extreme points of E are necessarily of the form $\alpha_i = 0$ for $i \notin \{a, b\}$ for some $a \neq b \in [n]$. Furthermore, since $\mathcal{L}$ is homogeneous in α, we may rescale it so that $\alpha_b = 1$. In this case we have

$$\alpha_a = \frac{A - f_b}{f_a}$$

and

$$B = \alpha_a g_a + g_b = \frac{A - f_b}{f_a} g_a + g_b.$$

Substituting this into (3.5), after some algebra we get

$$\mathcal{L} = \frac{(\beta_a - \beta_b)(A - f_b)(g_b f_a - g_a f_b)}{A(A g_a + g_b f_a - g_a f_b)}.$$

Maximizing $\mathcal{L}$ with respect to A yields

$$A^* = f_b + \sqrt{\frac{f_a f_b g_b}{g_a}},$$

with a maximal value of

$$\mathcal{L}(A^*) = (\beta_a - \beta_b) \frac{(g_b f_a - g_a f_b)\sqrt{f_a f_b g_b / g_a}}{(f_b + \sqrt{f_a f_b g_b / g_a})(g_b f_a + g_a \sqrt{f_a f_b g_b / g_a})}. \tag{3.6}$$

The nontrivial values for β_a, β_b are 1 and -1, respectively, and it remains to choose the $f_a, f_b, g_a, g_b \in \{r, R\}$ so as to maximize the above display. It is easily seen that the choice $f_a = g_b = R$ and $f_b = g_a = r$ is optimal, which yields the value

$$2\frac{(R^2 - r^2)R}{R(r + R)^2} = 2\frac{R - r}{R + r},$$

as claimed. □

Remark 3.2. In cases where the estimate $(R - r)/(R + r)$ is too crude, it might be possible to salvage a useful bound from (3.6).

Proof of Theorem 3.2. Let us define the shorthand notation:

$$\pi(u_k^l) = \prod_{t=k}^{l-1} \psi_{t,t+1}(u_t, u_{t+1}).$$

Then we expand

$$\begin{aligned}
p_i(x \mid y) = &\Bigg(\sum_{x' \in \Omega} \sum_{{v'}_1^{i-1}} \sum_{{z'}_{i+2}^n} \pi({v'}_1^{i-2}) \psi_{i-1,i}(v'_{i-1}, y) \psi_{i,i+1}(y, x') \\
&\times \psi_{i+1,i+2}(x', z'_{i+2}) \pi({z'}_{i+2}^n) \Bigg)^{-1} \\
&\times \sum_{v_1^{i-1}} \sum_{z_{i+2}^n} \pi(v_1^{i-2}) \psi_{i-1,i}(v_{i-1}, y) \psi_{i,i+1}(y, x) \psi_{i+1,i+2}(x, z_{i+2}) \pi(z_{i+2}^n) \\
= &\frac{\psi_{i,i+1}(y,x) a_{yx}}{\sum_{x' \in \Omega} \psi_{i,i+1}(y,x') a_{yx'}}
\end{aligned}$$

where

$$a_{yx} = \sum_{v_1^{i-1}} \sum_{z_{i+2}^n} \pi(v_1^{i-2}) \psi_{i-1,i}(v_{i-1}, y) \psi_{i+1,i+2}(x, z_{i+2}) \pi(z_{i+2}^n)$$

(we take the natural convention that $\psi_{i,j}(\cdot \mid \cdot) = 1$ whenever $(i,j) \notin E$).

Fix $y, y' \in \Omega$. Define the quantities, for each $x \in \Omega$:

$$\begin{aligned}
f_x &= \psi_{i,i+1}(y, x), \\
g_x &= \psi_{i,i+1}(y', x), \\
\alpha_x &= a_{yx}, \\
\alpha'_x &= a_{y'x}.
\end{aligned}$$

Then

$$\begin{aligned}
\sum_{x \in \Omega} \left| p_i(x \mid y) - p_i(x \mid y') \right| &= \sum_{x \in \Omega} \left| \frac{f_x \alpha_x}{\sum_{x' \in \Omega} f_{x'} \alpha_{x'}} - \frac{g_x \alpha'_x}{\sum_{x' \in \Omega} g_{x'} \alpha'_{x'}} \right| \\
&= \sum_{x \in \Omega} \left| \frac{f_x \alpha_x}{\sum_{x' \in \Omega} f_{x'} \alpha_{x'}} - \frac{g_x \alpha_x}{\sum_{x' \in \Omega} g_{x'} \alpha_{x'}} \right|;
\end{aligned}$$

the last equality follows since $\alpha'_x = c\alpha_x$, where $c = (\psi_{i-1,i}(v_{i-1}, y))^{-1} \times \psi_{i-1,i}(v_{i-1}, y')$. Now Lemma 3.1 can be applied to establish the claim. □

We observe that in general, mixing bounds on directed Markov chains are more informative than on undirected ones. In the language of random fields, the measure μ defined in (3.2) is a finite volume Gibbs measure with a pair potential. Let us recall the Dobrushin uniqueness condition and its role in concentration inequalities. For $j \in V$, define $\bar{V}_j := V \setminus \{j\}$ and $\sim_j$ to be the following equivalence relation on Ω^V: $x \sim_j y$ if $x_k = y_k$ for all $k \in \bar{V}_j$. Also, define the operator $\bar{(\cdot)}_i : \Omega^V \to \Omega^{\bar{V}_i}$ as the obvious coordinate projection. For $i \in V$ and $x \in \Omega^V$, define $\mu_i(\cdot \mid \bar{x}_j)$ to be the distribution on X_i conditioned on the other $\{X_j\}$ being equal to $\bar{x}_j$. Define the $n \times n$ *Dobrushin interdependence matrix* $D = (d_{ij})$:

$$d_{ij} = \max_{x \sim_j y \in \Omega^V} \|\mu_i(\cdot \mid \bar{x}_j) - \mu_i(\cdot \mid \bar{y}_j)\|. \tag{3.7}$$

The Dobrushin uniqueness condition requires that $\|D\|_\infty < 1$ (all norms on discussed here D are $\ell_p \to \ell_p$ operator norms). Let μ be any probability measure on Ω^V and for simplicity, take $f : \Omega^V \to \mathbb{R}$ to satisfy $\sqrt{n}\|f\|_{\mathrm{Lip}} \le 1$ with respect to the Hamming metric. Then a result of Külske [21, Theorem 1] states that

$$\mu\{|f - \mu f| \ge t\} \le 2 \exp\Big(- \frac{t^2}{2}(1 - \|D\|_\infty)(1 - \|D^\mathsf{T}\|_\infty)\Big), \quad t \ge 0, \tag{3.8}$$

provided that $\|D\|_\infty, \|D^\mathsf{T}\|_\infty < 1$, and Chatterjee in [6, Theorem 4.3] proves that

$$\mu\{|f - \mu f| \ge t\} \le 2 \exp\big(- (1 - \|D\|_2) t^2\big), \quad t \ge 0, \tag{3.9}$$

provided that $\|D\|_2 < 1$.

On the other hand, consider a directed homogeneous Markov chain $(X_1, \ldots, X_n)$ with measure μ on Ω^n induced by a Markov kernel with contraction coefficient $\theta < 1$. In this case, the ℓ_∞ norm of the η-mixing matrix (see 1.6) is easily bounded:

$$\|\Delta\|_\infty \le \frac{1}{1-\theta}.$$

An immediate consequence of the definition (3.7) is that $d_{n,n-1} = \theta$, which implies that

$$\theta \le \min\big\{\|D\|_\infty, \|D^\mathsf{T}\|_\infty, \|D\|_2\big\}. \tag{3.10}$$

Thus, for directed homogeneous Markov chains, (1.5) is uniformly superior to (3.8). The inequality (3.10) does not imply that (1.5) is uniformly superior to (3.9) — and indeed, this is not the case. However, (3.10) does imply that whenever (3.9) is nontrivial (i.e., $\|D\|_2 < 1$), the contraction bound (1.5) is also nontrivial (i.e., $\theta < 1$). This implication does not hold in the reverse, as we show by example. Take $\Omega = \{0, 1\}$ and consider the Markov kernel defined by $P(0 \mid 0) = a$, $P(0 \mid 1) = b$, $P(1 \mid 0) = 1 - a$, $P(1 \mid 1) = 1 - b$, for $a, b \in (0, 1)$.

A straightforward calculation shows that for $a \le 1/2$ and as $b \searrow 0$, we have $\theta \to a$ and

$$\min\left\{\|D\|_\infty, \|D^\mathsf{T}\|_\infty\right\} \longrightarrow 2/(1+a), \quad \|D\|_2 > 1.5/(1+a),$$

rendering the corresponding Dobrushin matrix-based estimates uninformative even for very small θ.

4. Markov tree processes

4.1. Preliminaries

We begin by defining some notation specific to this section. A collection of variables may be indexed by subset: if $x \in \Omega^V$ and $I \subseteq V$ with $I = \{i_1, i_2, \ldots, i_m\}$, then we write $x_I \equiv x[I] = \{x_{i_1}, x_{i_2}, \ldots, x_{i_m}\}$; we will write x_I and $x[I]$ interchangeably, as dictated by convenience. To avoid cumbersome subscripts, we will also occasionally use the bracket notation for vector components. Thus, if $\mathbf{u} \in \mathbb{R}^{\Omega^I}$, then

$$\mathbf{u}_{x_I} \equiv \mathbf{u}_{x[I]} \equiv \mathbf{u}[x_I] \equiv \mathbf{u}[x[I]] = \mathbf{u}_{(x_{i_1}, x_{i_2}, \ldots, x_{i_m})} \in \mathbb{R}$$

for each $x[I] \in \Omega^I$. A similar bracket notation will apply for matrices. If A is a matrix then $A_{*,j} = A[*, j]$ will denote its jth column. Probabilities are denoted by $\mathbf{P}$ in this section.

If $G = (V, E)$ is a graph, we will frequently abuse notation and write $u \in G$ instead of $u \in V$, blurring the distinction between a graph and its vertex set. This notation will carry over to set-theoretic operations ($G = G_1 \cap G_2$) and indexing of variables (e.g., X_G).

4.2. Graph theory

Consider a directed acyclic graph $G = (V, E)$, and define a partial order $\prec_G$ on G by the transitive closure of the relation

$$u \prec_G v \quad \text{if} \quad (u, v) \in E.$$

We define the *parents* and *children* of $v \in V$ in the natural way:

$$\text{parents}(v) = \{u \in V : (u, v) \in E\}$$

and

$$\text{children}(v) = \{w \in V : (v, w) \in E\}.$$

If G is connected and each $v \in V$ has at most one parent, G is called a (*directed*) *tree*. In a tree, whenever $u \prec_G v$ there is a unique directed path from u to v. A tree T always has a unique minimal (with respect to $\prec_T$) element

$r_0 \in V$, called its *root*. Thus, for every $v \in V$ there is a unique directed path $r_0 \prec_T r_1 \prec_T \ldots \prec_T r_d = v$; define the *depth* of v, $\mathrm{dep}_T(v) = d$, to be the length (i.e., number of edges) of this path. Note that $\mathrm{dep}_T(r_0) = 0$. We define the depth of the tree by $\mathrm{dep}(T) = \sup_{v \in T} \mathrm{dep}_T(v)$.

For $d = 0, 1, \ldots$ define the dth *level* of the tree T by

$$\mathrm{lev}_T(d) = \{v \in V : \mathrm{dep}_T(v) = d\};$$

note that the levels induce a disjoint partition on V:

$$V = \bigcup_{d=1}^{\mathrm{dep}(T)} \mathrm{lev}_T(d).$$

We define the *width*[2] of a tree as the greatest number of nodes in any level:

$$\mathrm{wid}(T) = \sup_{1 \le d \le \mathrm{dep}(T)} |\,\mathrm{lev}_T(d)|. \tag{4.1}$$

We will consistently take $|V| = n$ for finite V. An ordering $J : V \to \mathbb{N}$ of the nodes is said to be *breadth-first* if

$$\mathrm{dep}_T(u) < \mathrm{dep}_T(v) \Longrightarrow J(u) < J(v). \tag{4.2}$$

Since every finite directed tree $T = (V, E)$ has some breadth-first ordering,[3] we will henceforth blur the distinction between $v \in V$ and $J(v)$, simply taking $V = [n]$ (or $V = \mathbb{N}$) and assuming that $\mathrm{dep}_T(u) < \mathrm{dep}_T(v) \Rightarrow u < v$ holds. This will allow us to write Ω^V simply as Ω^n for any set Ω.

Note that we have two orders on V: the partial order $\prec_T$, induced by the tree edges, and the total order $<$, given by the breadth-first enumeration. Observe that $i \prec_T j$ implies $i < j$ but not vice versa.

If $T = (V, E)$ is a tree and $u \in V$, we define the *subtree* induced by u, $T_u = (V_u, E_u)$ by $V_u = \{v \in V : u \preceq_T v\}$, $E_u = \{(v, w) \in E : v, w \in V_u\}$.

4.3. Markov tree measure

If Ω is a finite set, a *Markov tree measure* μ is defined on Ω^n by a tree $T = (V, E)$ and transition kernels p_0, $\{p_{ij} : (i, j) \in E\}$. Continuing our convention above, we have a breadth-first order $<$ and the total order $\prec_T$ on V, and take $V = \{1, \ldots, n\}$. Together, the edges of T and the transition kernels determine the distribution μ on Ω^n:

$$\mu(x) = p_0(x_1) \prod_{(i,j) \in E} p_{ij}(x_j \mid x_i), \quad x \in \Omega^n. \tag{4.3}$$

[2]This definition is nonstandard.

[3]One can easily construct a breadth-first ordering on a given tree by ordering the nodes arbitrarily within each level and listing the levels in ascending order: $\mathrm{lev}_T(1), \mathrm{lev}_T(2), \ldots$

A measure on Ω^n satisfying (4.3) for some T and $\{p_{ij}\}$ is said to be *compatible* with tree T; a measure is a Markov tree measure if it is compatible with some tree.

Suppose Ω is a finite set and $(X_i)_{i\in\mathbb{N}}$, $X_i \in \Omega$ is a random process defined on $(\Omega^{\mathbb{N}}, \mathbf{P})$. If for each $n > 0$ there is a tree $T^{(n)} = ([n], E^{(n)})$ and a Markov tree measure μ_n compatible with $T^{(n)}$ such that for all $x \in \Omega^n$ we have

$$\mathbf{P}\{X_1^n = x\} = \mu_n(x)$$

then we call X a *Markov tree process*. The trees $\{T^{(n)}\}$ are easily seen to be consistent in the sense that $T^{(n)}$ is an induced subgraph of $T^{(n+1)}$. So corresponding to any Markov tree process is the unique infinite tree $T = (\mathbb{N}, E)$. The uniqueness of T is easy to see, since for $v > 1$, the parent of v is the smallest $u \in \mathbb{N}$ such that

$$\mathbf{P}\{X_v = x_v \mid X_1^u = x_1^u\} = \mathbf{P}\{X_v = x_v \mid X_u = x_u\};$$

thus $\mathbf{P}$ determines the edges of T.

It is straightforward to verify that a Markov tree process $\{X_v\}_{v\in T}$ compatible with tree T has the following *Markov property*: if v and v' are children of u in T, then

$$\begin{aligned}\mathbf{P}&\left\{X_{T_v} = x, X_{T_{v'}} = x' \mid X_u = y\right\}\\ &= \mathbf{P}\{X_{T_v} = x \mid X_u = y\}\mathbf{P}\{X_{T_{v'}} = x' \mid X_u = y\}.\end{aligned}$$

In other words, the subtrees induced by the children are conditionally independent given the parent; this follows directly from the definition of the Markov tree measure in (4.3).

4.4. Statement of result

Theorem 4.1. *Let Ω be a finite set and let $(X_i)_{1\le i\le n}$, $X_i \in \Omega$ be a Markov tree process, defined by tree $T = (V, E)$ and transition kernels p_0, $\{p_{uv}(\cdot \mid \cdot)\}_{(u,v)\in E}$. Define the (u, v)-contraction coefficient θ_{uv} by*

$$\theta_{uv} = \max_{y,y'\in\Omega} \|p_{uv}(\cdot \mid y) - p_{uv}(\cdot \mid y')\|. \tag{4.4}$$

Suppose $\max_{(u,v)\in E} \theta_{uv} \le \theta < 1$ for some θ and $\mathrm{wid}(T) \le L$. Then for the Markov tree process X we have

$$\bar{\eta}_{ij} \le \left(1 - (1-\theta)^L\right)^{\lfloor (j-i)/L \rfloor} \tag{4.5}$$

for $1 \le i < j \le n$.

To cast (4.5) in more usable form, we first note that for $k, L \in \mathbb{N}$ with $k \geq L$, we have

$$\left\lfloor \frac{k}{L} \right\rfloor \geq \frac{k}{2L-1} \tag{4.6}$$

(we omit the elementary number-theoretic proof). Using (4.6), we have

$$\bar{\eta}_{ij} \leq \tilde{\theta}^{j-i}, \quad \text{for } j \geq i + L \tag{4.7}$$

where

$$\tilde{\theta} = (1 - (1-\theta)^L)^{1/(2L-1)};$$

this implies the dimension-free bound

$$\|\Delta\|_\infty \leq L - 1 + (1 - \tilde{\theta})^{-1}.$$

In the (degenerate) case where the Markov tree is a chain, we have $L = 1$ and therefore $\tilde{\theta} = \theta$; thus we recover Theorem 3.2.

4.5. Proof of Theorem 4.1

The proof of Theorem 4.1 is combination of elementary graph theory and tensor algebra. We start with a graph-theoretic lemma:

Lemma 4.1. *Let $T = ([n], E)$ be a tree and fix $1 \leq i < j \leq n$. Suppose $(X_i)_{1 \leq i \leq n}$ is a Markov tree process whose distribution $\mathbf{P}$ on Ω^n is compatible with T (in the sense of Section 4.3). Define the set*

$$T_i^j = T_i \bigcap \{j, j+1, \ldots, n\},$$

consisting of those nodes in the subtree T_i whose breadth-first numbering does not precede j. Then, for $y \in \Omega^{i-1}$ and $w, w' \in \Omega$, we have

$$\eta_{ij}(y, w, w') = \begin{cases} 0, & T_i^j = \emptyset, \\ \eta_{ij_0}(y, w, w'), & \textit{otherwise}, \end{cases} \tag{4.8}$$

where j_0 is the minimum (with respect to $<$) element of T_i^j.

Remark 4.1. This lemma tells us that when computing η_{ij} it is sufficient to restrict our attention to the subtree induced by i.

Proof. The case $j \in T_i$ implies $j_0 = j$ and is trivial; thus we assume $j \notin T_i$. In this case, the subtrees T_i and T_j are disjoint. Putting $\bar{T}_i = T_i \setminus \{i\}$, we have by the Markov property,

$$\begin{aligned} &\mathbf{P}\left\{X_{\bar{T}_i} = x_{\bar{T}_i}, X_{T_j} = x_{T_j} \mid X_1^i = yw\right\} \\ &\quad = \mathbf{P}\left\{X_{\bar{T}_i} = x_{\bar{T}_i} \mid X_i = w\right\}\mathbf{P}\left\{X_{T_j} = x_{T_j} \mid X_1^{i-1} = y\right\}. \end{aligned}$$

Then from the definition of η_{ij} and by marginalizing out the X_{T_j}, we have

$$\begin{aligned}\eta_{ij}(y,w,w') &= \frac{1}{2}\sum_{x_j^n} \left|\mathbf{P}\left\{X_j^n = x_j^n \mid X_1^i = yw\right\} - \mathbf{P}\left\{X_j^n = x_j^n \mid X_1^i = yw'\right\}\right| \\ &= \frac{1}{2}\sum_{x_{T_i^j}} \left|\mathbf{P}\left\{X_{T_i^j} = x_{T_i^j} \mid X_i = w\right\} - \mathbf{P}\left\{X_{T_i^j} = x_{T_i^j} \mid X_i = w'\right\}\right|.\end{aligned}$$

If $T_i^j = \emptyset$ then obviously $\eta_{ij} = 0$; otherwise, $\eta_{ij} = \eta_{ij_0}$, since j_0 is the "first" element of T_i^j. □

Next we develop some basic results for tensor norms. If $\mathbf{A}$ is an $M \times N$ column-stochastic matrix (i.e., $\mathbf{A}_{ij} \geq 0$ for $1 \leq i \leq M$, $1 \leq j \leq N$ and $\sum_{i=1}^M \mathbf{A}_{ij} = 1$ for all $1 \leq j \leq N$) and $\mathbf{u} \in \mathbb{R}^N$ is *balanced* in the sense that $\sum_{j=1}^N \mathbf{u}_j = 0$, we have, by Lemma 2.1

$$\|\mathbf{Au}\| \leq \|\mathbf{A}\|\|\mathbf{u}\|, \tag{4.9}$$

where

$$\|\mathbf{A}\| = \max_{1\leq j,j'\leq N} \|\mathbf{A}_{*,j} - \mathbf{A}_{*,j'}\|, \tag{4.10}$$

and $\mathbf{A}_{*,j} \equiv \mathbf{A}[*,j]$ denotes the jth column of $\mathbf{A}$. An immediate consequence of (4.9) is that $\|\cdot\|$ satisfies

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\| \tag{4.11}$$

for column-stochastic matrices $\mathbf{A} \in \mathbb{R}^{M\times N}$ and $\mathbf{B} \in \mathbb{R}^{N\times P}$.

Remark 4.2. Note that if $\mathbf{A}$ is a column-stochastic matrix then $\|\mathbf{A}\| \leq 1$, and if additionally $\mathbf{u}$ is balanced then $\mathbf{Au}$ is also balanced.

If $\mathbf{u} \in \mathbb{R}^M$ and $\mathbf{v} \in \mathbb{R}^N$, define their tensor product $\mathbf{w} = \mathbf{v} \otimes \mathbf{u}$ by

$$\mathbf{w}_{(i,j)} = \mathbf{u}_i \mathbf{v}_j,$$

where the notation $(\mathbf{v} \otimes \mathbf{u})_{(i,j)}$ is used to distinguish the 2-tensor $\mathbf{w}$ from an $M \times N$ matrix. The tensor $\mathbf{w}$ is a vector in $\mathbb{R}^{MN}$ indexed by pairs $(i,j) \in [M] \times [N]$; its norm is naturally defined to be

$$\|\mathbf{w}\| = \frac{1}{2}\sum_{(i,j)\in[M]\times[N]} |\mathbf{w}_{(i,j)}|. \tag{4.12}$$

To develop a convenient tensor notation, we will fix the index set $V = \{1, \ldots, n\}$. For $I \subset V$, a tensor indexed by I is a vector $\vec{u} \in \mathbb{R}^{\Omega^I}$. A special case of such an I-tensor is the product $\mathbf{u} = \bigotimes_{i \in I} \mathbf{v}^{(i)}$, where $\mathbf{v}^{(i)} \in \mathbb{R}^{\Omega}$ and

$$\mathbf{u}[x_I] = \prod_{i \in I} \mathbf{v}^{(i)}[x_i]$$

for each $x_I \in \Omega^I$. To gain more familiarity with the notation, let us write the total variation norm of an I-tensor:

$$\|\mathbf{u}\| = \frac{1}{2} \sum_{x_I \in \Omega^I} \big|\mathbf{u}[x_I]\big|. \tag{4.13}$$

In order to extend Lemma 2.2 to product tensors, we will need to define the function $\alpha_k : \mathbb{R}^k \to \mathbb{R}$ and state some of its properties:

Lemma 4.2. *Define $\alpha_k : \mathbb{R}^k \to \mathbb{R}$ recursively as $\alpha_1(x) = x$ and*

$$\alpha_{k+1}(x_1, x_2, \ldots, x_{k+1}) = x_{k+1} + (1 - x_{k+1})\alpha_k(x_1, x_2, \ldots, x_k). \tag{4.14}$$

Then

(a) *α_k is symmetric in its k arguments, so it is well-defined as a mapping*

$$\alpha : \{x_i : 1 \le i \le k\} \mapsto \mathbb{R}$$

from finite real sets to the reals;

(b) *α_k takes $[0,1]^k$ to $[0,1]$ and is monotonically increasing in each argument on $[0,1]^k$;*

(c) *if $B \subset C \subset [0,1]$ are finite sets then $\alpha(B) \le \alpha(C)$;*

(d) *$\alpha_k(x, x, \ldots, x) = 1 - (1 - x)^k$;*

(e) *if B is finite and $1 \in B \subset [0,1]$ then $\alpha(B) = 1$;*

(f) *if $B \subset [0,1]$ is a finite set then $\alpha(B) \le \sum_{x \in B} x$.*

Remark 4.3. In light of (a), we will use the notation $\alpha_k(x_1, x_2, \ldots, x_k)$ and $\alpha(\{x_i : 1 \le i \le k\})$ interchangeably, as dictated by convenience.

Proof. Claims (a), (b), (e), (f) are straightforward to verify from the recursive definition of α and induction. Claim (c) follows from (b) since

$$\alpha_{k+1}(x_1, x_2, \ldots, x_k, 0) = \alpha_k(x_1, x_2, \ldots, x_k)$$

and (d) is easily derived from the binomial expansion of $(1 - x)^k$. □

The function α_k is the natural generalization of $\alpha_2(x_1, x_2) = x_1 + x_2 - x_1 x_2$ to k variables, and it is what we need for the analog of Lemma 2.2 for a product of k tensors:

Corollary 4.1. *Let $\{\mathbf{u}^{(i)}\}_{i\in I}$ and $\{\mathbf{v}^{(i)}\}_{i\in I}$ be two sets of tensors and assume that each of $\mathbf{u}^{(i)}$, $\mathbf{v}^{(i)}$ is a probability measure on Ω. Then we have*

$$\Big\| \bigotimes_{i\in I} \mathbf{u}^{(i)} - \bigotimes_{i\in I} \mathbf{v}^{(i)} \Big\| \le \alpha\big\{ \|\mathbf{u}^{(i)} - \mathbf{v}^{(i)}\| : i \in I \big\}. \tag{4.15}$$

Proof. Pick an $i_0 \in I$ and let $\mathbf{p} = \mathbf{u}^{(i_0)}$, $\mathbf{q} = \mathbf{v}^{(i_0)}$,

$$\mathbf{p}' = \bigotimes_{i_0 \ne i \in I} \mathbf{u}^{(i)}, \quad \mathbf{q}' = \bigotimes_{i_0 \ne i \in I} \mathbf{v}^{(i)}.$$

Apply Lemma 2.2 to $\|\mathbf{p} \otimes \mathbf{q} - \mathbf{p}' \otimes \mathbf{q}'\|$ and proceed by induction. □

Our final generalization concerns linear operators over I-tensors. For I, $J \subseteq V$, an I, J-matrix $\mathbf{A}$ has dimensions $|\Omega^J| \times |\Omega^I|$ and takes an I-tensor $\mathbf{u}$ to a J-tensor $\mathbf{v}$: for each $y_J \in \Omega^J$, we have

$$\mathbf{v}[y_J] = \sum_{x_I \in \Omega^I} \mathbf{A}[y_J, x_I]\mathbf{u}[x_I], \tag{4.16}$$

which we write as $\mathbf{A}\mathbf{u} = \mathbf{v}$. If $\mathbf{A}$ is an I, J-matrix and $\mathbf{B}$ is a J, K-matrix, the matrix product $\mathbf{B}\mathbf{A}$ is defined analogously to (4.16).

As a special case, an I, J-matrix might factorize as a tensor product of $|\Omega| \times |\Omega|$ matrices $\mathbf{A}^{(i,j)} \in \mathbb{R}^{\Omega\times\Omega}$. We will write such a factorization in terms of a bipartite graph[4] $G = (I + J, E)$, where $E \subset I \times J$ and the factors $\mathbf{A}^{(i,j)}$ are indexed by $(i, j) \in E$:

$$\mathbf{A} = \bigotimes_{(i,j)\in E} \mathbf{A}^{(i,j)}, \tag{4.17}$$

where

$$\mathbf{A}[y_J, x_I] = \prod_{(i,j)\in E} \mathbf{A}^{(i,j)}_{y_j, x_i}$$

for all $x_I \in \Omega^I$ and $y_J \in \Omega^J$. The norm of an I, J-matrix is a natural generalization of the matrix norm defined in (4.10):

$$\|\mathbf{A}\| = \max_{x_I, x'_I \in \Omega^I} \big\| \mathbf{A}[*, x_I] - \mathbf{A}[*, x'_I] \big\| \tag{4.18}$$

[4]Our notation for bipartite graphs is standard; it is equivalent to $G = (I \cup J, E)$ where I and J are always assumed to be disjoint.

where $\mathbf{u} = \mathbf{A}[*, x_I]$ is the J-tensor given by

$$\mathbf{u}[y_J] = \mathbf{A}[y_J, x_I];$$

(4.18) is well-defined via the tensor norm in (4.13). Since I, J matrices act on I-tensors by ordinary matrix multiplication, $\|\mathbf{Au}\| \le \|\mathbf{A}\|\|\mathbf{u}\|$ continues to hold when $\mathbf{A}$ is a column-stochastic I, J-matrix and $\mathbf{u}$ is a balanced I-tensor; if, additionally, $\mathbf{B}$ is a column-stochastic J, K-matrix, $\|\mathbf{BA}\| \le \|\mathbf{B}\|\|\mathbf{A}\|$ also holds. Likewise, since another way of writing (4.17) is

$$\mathbf{A}[*, x_I] = \bigotimes_{(i,j)\in E} \mathbf{A}^{(i,j)}[*, x_i],$$

Corollary 4.1 extends to tensor products of matrices:

Lemma 4.3. *Fix index sets I, J and a bipartite graph $(I + J, E)$. Let $\{\mathbf{A}^{(i,j)}\}_{(i,j)\in E}$ be a collection of column-stochastic $|\Omega| \times |\Omega|$ matrices, whose tensor product is the I, J matrix*

$$\mathbf{A} = \bigotimes_{(i,j)\in E} \mathbf{A}^{(i,j)}.$$

Then

$$\|\mathbf{A}\| \le \alpha\big\{\|\mathbf{A}^{(i,j)}\| : (i,j) \in E\big\}.$$

We are now in a position to state the main technical lemma, from which Theorem 4.1 will follow straightforwardly:

Lemma 4.4. *Let Ω be a finite set and let $(X_i)_{1\le i\le n}$, $X_i \in \Omega$, be a Markov tree process, defined by a tree $T = (V, E)$ and transition kernels*

$$p_0, \ \{p_{uv}(\cdot \mid \cdot)\}_{(u,v)\in E}.$$

Let the (u, v)-contraction coefficient θ_{uv} be as defined in (4.4).

Fix $1 \le i < j \le n$ and let $j_0 = j_0(i, j)$ be as defined in Lemma 4.1 (we are assuming its existence, for otherwise $\bar\eta_{ij} = 0$). Then we have

$$\bar\eta_{ij} \le \prod_{d=\mathrm{dep}_T(i)+1}^{\mathrm{dep}_T(j_0)} \alpha\{\theta_{uv} : v \in \mathrm{lev}_T(d)\}. \tag{4.19}$$

Proof. For $y \in \Omega^{i-1}$ and $w, w' \in \Omega$, we have

$$\begin{aligned}\eta_{ij}(y, w, w') &= \frac{1}{2}\sum_{x_j^n} \big|\mathbf{P}\{X_j^n = x_j^n \mid X_1^i = yw\} - \mathbf{P}\{X_j^n = x_j^n \mid X_1^i = yw'\}\big| \\ &= \frac{1}{2}\sum_{x_j^n} \Big| \sum_{z_{i+1}^{j-1}} \big(\mathbf{P}\{X_{i+1}^n = z_{i+1}^{j-1}\, x_j^n \mid X_1^i = yw\} \\ &\qquad - \mathbf{P}\{X_{i+1}^n = z_{i+1}^{j-1}\, x_j^n \mid X_1^i = yw'\}\big)\Big|. \end{aligned} \tag{4.20}$$

Let T_i be the subtree induced by i and

$$Z = T_i \bigcap \{i+1, \dots, j_0 - 1\} \quad \text{and} \quad C = \{v \in T_i : (u,v) \in E, u < j_0, v \geq j_0\}. \tag{4.21}$$

Then by Lemma 4.1 and the Markov property, we get

$$\eta_{ij}(y, w, w') = \frac{1}{2} \sum_{x[C]} \Big| \sum_{x[Z]} \big(\mathbf{P}\{X[C \cup Z] = x[C \cup Z] \mid X_i = w\} \tag{4.22}$$
$$- \mathbf{P}\{X[C \cup Z] = x[C \cup Z] \mid X_i = w'\} \big) \Big|$$

(the sum indexed by $\{j_0, \dots, n\} \setminus C$ marginalizes out).

Define $D = \{d_k : k = 0, \dots, |D|\}$ with $d_0 = \operatorname{dep}_T(i)$, $d_{|D|} = \operatorname{dep}_T(j_0)$ and $d_{k+1} = d_k + 1$ for $0 \leq k < |D|$. For $d \in D$, let $I_d = T_i \cap \operatorname{lev}_T(d)$ and $G_d = (I_{d-1} + I_d, E_d)$ be the bipartite graph consisting of the nodes in I_{d-1} and I_d, and the edges in E joining them (note that $I_{d_0} = \{i\}$).

For $(u, v) \in E$, let $\mathbf{A}^{(u,v)}$ be the $|\Omega| \times |\Omega|$ matrix given by

$$\mathbf{A}^{(u,v)}_{x,x'} = p_{uv}(x \mid x')$$

and note that $\|\mathbf{A}^{(u,v)}\| = \theta_{uv}$. Then by the Markov property, for each $z[I_d] \in \Omega^{I_d}$ and $x[I_{d-1}] \in \Omega^{I_{d-1}}$, $d \in D \setminus \{d_0\}$, we have

$$\mathbf{P}\{X_{I_d} = z_{I_d} \mid X_{I_{d-1}} = x_{I_{d-1}}\} = \mathbf{A}^{(d)}[z_{I_d}, x_{I_{d-1}}],$$

where

$$\mathbf{A}^{(d)} = \bigotimes_{(u,v) \in E_d} \mathbf{A}^{(u,v)}.$$

Likewise, for $d \in D \setminus \{d_0\}$,

$$\begin{aligned} &\mathbf{P}\left\{X_{I_d} = x_{I_d} \mid X_i = w\right\} \\ &\quad = \sum_{x'_{I_1}} \sum_{x''_{I_2}} \cdots \sum_{x^{(d-1)}_{I_{d-1}}} \mathbf{P}\{X_{I_1} = x'_{I_1} \mid X_i = w\} \\ &\qquad \times \mathbf{P}\{X_{I_2} = x''_{I_2} \mid X_{I_1} = x'_{I_1}\} \cdots \mathbf{P}\{X_{I_d} = x_{I_d} \mid X_{I_{d-1}} = x^{(d-1)}_{I_{d-1}}\} \\ &\quad = \big(\mathbf{A}^{(d)} \mathbf{A}^{(d-1)} \cdots \mathbf{A}^{(d_1)}\big)[x_{I_d}, w]. \end{aligned} \tag{4.23}$$

Define the (balanced) I_{d_1}-tensor

$$\mathbf{h} = \mathbf{A}^{(d_1)}[*, w] - \mathbf{A}^{(d_1)}[*, w'], \tag{4.24}$$

the $I_{d_{|D|}}$-tensor

$$\mathbf{f} = \mathbf{A}^{(d_{|D|})} \mathbf{A}^{(d_{|D|-1})} \cdots \mathbf{A}^{(d_2)} \mathbf{h}, \tag{4.25}$$

and $C_0, C_1, Z_0 \subset \{1, \ldots, n\}$:

$$C_0 = C \cap I_{\mathrm{dep}_T(j_0)}, \quad C_1 = C \setminus C_0, \quad Z_0 = I_{\mathrm{dep}_T(j_0)} \setminus C_0, \tag{4.26}$$

where C and Z are defined in (4.21). For readability we will write $\mathbf{P}(x_U \mid \cdot)$ instead of $\mathbf{P}\{X_U = x_U \mid \cdot\}$ below; no ambiguity should arise. Combining (4.22) and (4.23), we have

$$\begin{aligned}\eta_{ij}(y, w, w') &= \frac{1}{2} \sum_{x_C} \Big| \sum_{x_Z} \big(\mathbf{P}(x[C \cup Z] \mid X_i = w) - \mathbf{P}(x[C \cup Z] \mid X_i = w') \big) \Big| \\ &= \frac{1}{2} \sum_{x_{C_0}} \sum_{x_{C_1}} \Big| \sum_{x_{Z_0}} \mathbf{P}(x[C_1] \mid x[Z_0]) \mathbf{f}[C_0 \cup Z_0] \Big| = \|\mathbf{B}\mathbf{f}\|\end{aligned}$$

where $\mathbf{B}$ is the $|\Omega^{C_0 \cup C_1}| \times |\Omega^{C_0 \cup Z_0}|$ column-stochastic matrix given by

$$\mathbf{B}[x_{C_0} \cup x_{C_1}, x'_{C_0} \cup x_{Z_0}] = \mathbb{1}\{x_{C_0} = x'_{C_0}\} \mathbf{P}(x_{C_1} \mid x_{Z_0})$$

with the convention that $\mathbf{P}(x_{C_1} \mid x_{Z_0}) = 1$ if either of Z_0 or C_1 is empty. The claim now follows by reading off the results previously obtained:

$$\begin{aligned}\|\mathbf{B}\mathbf{f} &\le \|\mathbf{B}\| \|\mathbf{f}\| && \text{Eq. (4.9)} \\ &\le \|\mathbf{f}\| && \text{Remark 4.2} \\ &\le \|\mathbf{h}\| \prod_{k=2}^{|D|} \|\mathbf{A}^{(d_k)}\| && \text{Eqs. (4.11), (4.25)} \\ &\le \prod_{k=1}^{|D|} \alpha\big\{ \|\mathbf{A}^{(u,v)}\| : (u,v) \in E_{d_k} \big\} && \text{Lemma 4.3.}\end{aligned}$$

□

Proof of Theorem 4.1. We will borrow the definitions from the proof of Lemma 4.4. To upper-bound $\bar{\eta}_{ij}$ we first bound $\alpha\{\|\mathbf{A}^{(u,v)}\| : (u,v) \in E_{d_k}\}$. Since

$$|E_{d_k}| \le \mathrm{wid}(T) \le L$$

(because every node in I_{d_k} has exactly one parent in $I_{d_{k-1}}$) and

$$\|\mathbf{A}^{(u,v)}\| = \theta_{uv} \le \theta < 1,$$

we appeal to Lemma 4.2 to obtain

$$\alpha\{\|\mathbf{A}^{(u,v)}\| : (u,v) \in E_{d_k}\} \le 1 - (1 - \theta)^L. \tag{4.27}$$

Now we must lower-bound the quantity $h = \mathrm{dep}_T(j_0) - \mathrm{dep}_T(i)$. Since every level can have up to L nodes, we have

$$j_0 - i \le hL$$

and so $h \ge \lfloor (j_0 - i)/L \rfloor \ge \lfloor (j - i)/L \rfloor$. □

The calculations in Lemma 4.4 yield considerably more information than the simple bound in (4.5) — certainly, the estimate in (4.19) is quite a bit sharper. Furthermore, suppose the tree T has levels $\{I_d : d = 0, 1, \ldots\}$ with the property that the levels are growing at most linearly:

$$|I_d| \leq cd$$

for some $c > 0$. Let $d_i = \mathrm{dep}_T(i)$, $d_j = \mathrm{dep}_T(j_0)$, and $h = d_j - d_i$. Then

$$\begin{aligned} j - i \leq j_0 - i \leq c \sum_{d_i+1}^{d_j} k &= \frac{c}{2}(d_j(d_j+1) - d_i(d_i+1)) \\ &< \frac{c}{2}((d_j+1)^2 - d_i^2) < \frac{c}{2}(d_i + h + 1)^2, \end{aligned}$$

so

$$h > \sqrt{2(j-i)/c} - d_i - 1,$$

which yields the bound, via Lemma 4.2 (f),

$$\bar{\eta}_{ij} \leq \prod_{k=1}^{h} \sum_{(u,v)\in E_k} \theta_{uv}. \tag{4.28}$$

Let $\theta_k = \max\{\theta_{uv} : (u,v) \in E_k\}$; then if $ck\theta_k \leq \beta$ holds for some $\beta \in \mathbb{R}$, this becomes

$$\bar{\eta}_{ij} \leq \prod_{k=1}^{h} (ck\theta_k) < \prod_{k=1}^{\sqrt{2(j-i)/c}-d_i-1} (ck\theta_k) \leq \beta^{\sqrt{2(j-i)/c}-d_i-1}. \tag{4.29}$$

This is a non-trivial bound for trees with linearly growing levels: recall that to bound $\|\Delta\|_\infty$, we must bound the series

$$\sum_{j=i+1}^{\infty} \bar{\eta}_{ij}.$$

By the limit comparison test with the series $\sum_{j=1}^{\infty} 1/j^2$, we have that

$$\sum_{j=i+1}^{\infty} \beta^{\sqrt{2(j-i)/c}-d_i-1}$$

converges for $\beta < 1$. Similar techniques may be applied when the level growth is bounded by other slowly increasing functions. It is hoped that this method will be extended to obtain concentration bounds for larger classes of directed acyclic graphical models.

Acknowledgments

Many thanks to Gideon Schechtman for hosting and guidance at the Weizmann Institute, and to Katalin Marton, Elchanan Mossel, and Cosma Shalizi for helpful comments on the manuscript. Thanks to Roi Weiss and an anonymous referee for insisting on a correct proof of Lemma 3.1.

References

[1] Y.F. Atchadé and J.S. Rosenthal (2005) On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* **11** (5), 815–828.

[2] K. Azuma (1967) Weighted sums of certain dependent random variables. *Tohoku Math. J.* **19**, 357–367.

[3] F. Barthe and A.V. Kolesnikov (2008) Mass transport and variants of the logarithmic Sobolev inequality. *J. Geom. Anal.* **18** (4), 921–979.

[4] I. Benjamini and Y. Peres (1994) Markov chains indexed by trees. *Ann. Prob.* **22** (1), 219–243.

[5] R.C. Bradley (2005) Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv. 2*, 107–144.

[6] S. Chatterjee (2005) Concentration inequalities with exchangeable pairs. Ph. D. Thesis, Stanford University.

[7] J.-R. Chazottes, P. Collet, C. Külske and F. Redig (2007) Concentration inequalities for random fields via coupling. *Probab. Theory and Relat. Fields* **137** (1–2), 201–225.

[8] J.-R. Chazottes and F. Redig (2009) Concentration inequalities for Markov processes via coupling. *Electron. J. Probab.* **14** (40), 1162–1180.

[9] H. Djellout, M.A. Guillin and L.M. Wu (2004) Transportation cost-information inequalities for random dynamical systems and diffusions. *Ann. Prob.* **32** (3B), 2702–2732.

[10] R. Dobrushin (1956) Central limit theorem for nonstationary Markov chains. II. *Teor. Verojatnost. i Primenen.* **1**, 365–425.

[11] P. Doukhan and S. Louhichi (1999) A new weak dependence condition and applications to moment inequalities. *Stoch. Process. Appl.* **84** (2), 313–342.

[12] D. Gamarnik (2003) Extension of the PAC framework to finite and countable Markov chains. *IEEE Trans. Inform. Theory* **49** (1), 338–345.

[13] A. Guillin, C. Léonard, L. Wu and N. Yao (2009) Transportation-information inequalities for Markov processes. *Probab. Theory and Relat. Fields* **144** (3), 669–695.

[14] W. Hoeffding (1963) Probability inequalities for sums of bounded random variables. *Amer. Statist. Assoc. J.* **58**, 13–30.

[15] R.A. Horn and C.R. Johnson (1985) *Matrix Analysis.* Cambridge University Press.

[16] A. Joulin and Y. Ollivier (2010) Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Prob.* **38** (6), 2418–2442.

[17] A. Kolesnikov (2010) Transportnaya zadacha i kontsentratsia (in Russian). In: *Introduction to the Mathematical Modeling of Transport Flows*, A.B. Gasnikov (ed.), MFTI, 288–298.

[18] L. Kontorovich (2007) Measure concentration of strongly mixing processes with applications. Ph. D. Thesis, Carnegie Mellon University.

[19] L. Kontorovich and A.E. Brockwell (2012) A strong law of large numbers for strongly mixing processes. To appear in: *Communications in Statistics — Theory and Methods.*

[20] L. Kontorovich and K. Ramanan (2008) Concentration inequalities for dependent random variables via the martingale method. *Ann. Prob.* **36** (6), 2126–2158.

[21] C. Külske (2003) Concentration inequalities for functions of Gibbs fields with application to diffraction and random Gibbs measures. *Commun. Math. Phys.* **239**, 29–51.

[22] S. Lauritzen (1996) *Graphical Models.* Oxford University Press.

[23] C. Manning and H. Schütze (1999) *Foundations of Statistical Natural Language Processing.* MIT Press.

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

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

[26] K. Marton (1998) Measure concentration for a class of random processes. *Probab. Theory and Relat. Fields* **110** (3), 427–439.

[27] K. Marton (2003) Measure concentration and strong mixing. *Stud. Sci. Math. Hungarica* **19** (1–2), 95–113.

[28] K. Marton (2004) Measure concentration for Euclidean distance in the case of dependent random variables. *Ann. Prob.* **32** (3), 2526–2544.

[29] P.S. Maybeck (1979) *Stochastic Models, Estimation, and Control.* Mathematics in Science and Engineering **141**.

[30] C. McDiarmid (1989) On the method of bounded differences. In: *Surveys in Combinatorics*, J. Siemons (ed.), LMS Lecture Notes Series **141**, Morgan Kaufmann Publishers, San Mateo, CA, 148–188.

[31] Y. Ollivier (2009) Ricci curvature of Markov chains on metric spaces. *J. Func. Anal.* **256** (3), 810–864.

[32] R. Pemantle (1995) Tree-indexed processes. *Statist. Sci.* **10** (2), 200–213.

[33] L.R. Rabiner (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77** (2), 257–286.

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

[35] P.-M. Samson (2000) Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Prob.* **28** (1), 416–461.