CUTOFF FOR A CLASS OF AUTO-REGRESSIVE MODELS WITH LARGE INITIALIZATION

BALÁZS GERENCSÉR AND ANDREA OTTOLINI

Abstract. We analyze the convergence rate of auto-regressive Markov chains \((X_k)_{k \geq 0}\) on \(\mathbb{R}^d\), where at each step a randomly chosen coordinate is replaced by a noisy damped weighted average of the others. The interest in the model comes from the connection with a certain Bayesian scheme utilized in the analysis of partially exchangeable data. Our main result shows that, under mild assumptions on the noise, a cutoff phenomenon occurs as the initialization \(X_0\) becomes large.

1. Introduction

Markov chains are used on a daily basis to sample from intractable distributions [8]. Under suitable ergodicity assumptions one is guaranteed that, after many iterations, a sample from the chain resembles that of its stationary distribution. For both practitioners and theoreticians, a natural question is to understand what “many” and “resemble” mean in this context.

Our interest will be in a class of measures on \(\mathbb{R}^d\) for some \(d \geq 2\). A classical way to approach the problem goes as follows: if \(\pi_k\) denotes the law of the Markov chain after \(k\) steps, and if \(\pi\) denotes its stationary measure, one is trying to understand how the total variation distance to stationarity

\[
\sup_{E \subset \mathbb{R}^d} |\pi_k(E) - \pi(E)| = \inf_{X_k \sim \pi_k, X \sim \pi} P(X_k \neq X),
\]

varies as \(k\) increases, the last inequality being the well-known coupling interpretation of total variation distance. In the display above, the supremum is taken over all Borel sets while the infimum is taken over all couplings of \(\pi\) and \(\pi_k\).

Often, the evolution of the chain depends on an additional parameter \(n\) (e.g., the size of the state space or some norm of the initial condition) and then it becomes important to understand what is the right sequence \(k = k(n)\) at which the transition to randomness occurs, i.e., the total variation distance drops as needed. For a friendly introduction to the slew of techniques and results on the subject the reader can refer to [11].

We consider Markov chains \((X_k)\) on \(\mathbb{R}^d\) that updates coordinates one at a time according to the auto-regressive scheme [13]. The regime of interest is that of a small additive noise (or, equivalently, large initial conditions). Informally, our result is that under mild assumptions and for large initial conditions \(\|X_0\| = n\), the chain takes about \(\ln n\) steps to mix. Moreover, we also prove that the transition to randomness occurs in a window of size \(\sqrt{\ln n}\). This is referred to as the cutoff phenomenon [4]. We also determine the location of the cutoff – i.e., the constant factor of the leading term \(\ln n\) – which is closely related to the convergence of a certain auxiliary Markov chain on the unit sphere.

Date: September 7, 2022.
We now proceed with a formal definition of the model.

1.1. The setup. Given $d \geq 2$, let $P = (p_{ij})_{1 \leq i, j \leq d}$ be the transition probabilities of a connected network without loops. For $x \in \mathbb{R}^d$, we define $\hat{x}$ by

$$\hat{x}_i := (Px)_i = \sum_{j=1}^{d} p_{ij} x_j,$$

(1.1)

where the sum actually runs over $j \neq i$ owing to the assumption that the network has no loops. Given $\epsilon_1, \ldots, \epsilon_d \in (0, 1)$, define $A_i : \mathbb{R}^d \to \mathbb{R}^d$, for $1 \leq i \leq d$, by setting

$$(A_i x)_j = x_j, \quad (A_i x)_i = \epsilon_i \hat{x}_i.$$  

(1.2)

Also, given $\sigma_1, \ldots, \sigma_d \in (0, \infty)$, define $b_i : \mathbb{R} \to \mathbb{R}^d$ by

$$(b_i(z))_j = 0, \quad (b_i(z))_i = \sigma_i z.$$  

(1.3)

Let $U$ denote the uniform measure on $\{1, \ldots, d\}$, and let $\gamma$ be an absolutely continuous probability measure on $\mathbb{R}$ with $\{\max(\log x, 0) \gamma(dx) < \infty \}$. Given $X_0 \in \mathbb{R}^d$ and independent random variables $I_1, Z_1, I_2, Z_2, \ldots$ where the $I_i$’s are distributed according to $U$, while the $Z_i$’s are distributed according to $\gamma$, define now a Markov chain on $\mathbb{R}^d$ via

$$X_k = A_{I_k} X_{k-1} + b_{I_k}(Z_k).$$

(1.4)

Owing to the assumption of $\gamma$, we are guaranteed by Theorem 2.1 in [5] that $X_k$ has a unique stationary distribution, which is the law of $X$ defined in terms of the backward iteration

$$X = b_{I_1}(Z_1) + A_{I_1} b_{I_2}(Z_2) + A_{I_1} A_{I_2} b_{I_3}(Z_3) + \ldots$$

(1.5)

Our main goal is to analyze the rate of convergence to stationarity for a large class of initial data.

**Theorem 1.1.** Let $\pi_k$ and $\pi$ be the laws of $X_k$ and $X$ as defined above. Consider a sequence of initial conditions $(X_0(n))_{n \geq 1}$ with positive coordinates and $\|X_0(n)\| = n$, and assume that the ratio between the minimum and maximum coordinate is bounded away from 0 uniformly in $n$.

Then, there exists a constant $\alpha \in (-\infty, 0)$ independent of $n$ such that, if

$$k = k(n, \beta) := \frac{\ln n + \beta \sqrt{\ln n}}{-\alpha},$$

(1.6)

then we have

$$\lim_{\beta \to \infty} \lim_{n \to \infty} d_{tv}(\pi, \pi_k) = 1, \quad \lim_{\beta \to \infty} \lim_{n \to \infty} d_{tv}(\pi, \pi_k) = 0.$$

**Remark 1.2.** The constant $\alpha$ is defined in terms of a certain auxiliary Markov chain on the unit sphere (see (1.5)), though its explicit value is inaccessible in general.

**Remark 1.3.** As it will be clear from the proof, the second conclusion of the theorem – i.e. the limit as $\beta \to +\infty$ – holds even for sequences with a coordinate ratio approaching to zero. On the other hand, the first conclusion does not hold in the case $X_0(n) = (n, 0, \ldots, 0)$ since, with positive probability (namely, if the first coordinate is selected first) the chain will mix in a bounded number of steps. If $X_0(n)$ has non-negative coordinates with at least two of them being strictly positive, it is easy to show that with high probability all coordinates will be of order $n$ in a bounded number of steps, and thus our result applies.
Let us give an overview of the main heuristic behind the proof. We start by analyzing the chain that we obtain by averaging over the randomness stemming from the $Z_k$'s. The core of the proof is to show that this chain is $O(1)$ with high probability precisely when $k$ is given by (1.5) for some fixed $\beta$. Then, the concentration properties of the stationary distribution and the absolute continuity of $\gamma$ allow us to conclude.

In the case $d = 2$, re-sampling the same coordinate has no effect on the distribution of the Markov chain, so that one can think of choosing coordinates in a deterministic fashion. If $\gamma$ is the law of a normal random variable, this allows numerical estimation of the total variation distance, displayed below, in striking accordance with the theoretical results.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Numerical estimation of the total variation distance between $\pi_k$ and $\pi$, in the case $d = 2$, $e_1 = e_2 = 0.55$, and $\gamma$ being a normal random variable. This results in $\alpha \approx -0.588$. The curves are drawn for ten values of $n$ in a geometric progression with ratio 5, starting from $n = 1000$. Both the logarithmic scaling and the cutoff are visible. Moreover, the horizontal distance between the curves is approximately $\ln 5 \approx 2.738$, as predicted by our theorem.}
\end{figure}

In general, to estimate the total variation distance one needs a more careful approach, even when $\gamma$ is the law of a normal random variable. Indeed, one has to approximate the distance between mixtures of normal random variables, for which no explicit formulas are available. However, our Theorem 1 guarantees that both the logarithmic scaling and the cutoff are extremely robust.

1.2. Structure of the paper. The rest of the paper is organized as follows. In Section 2 we briefly review a statistical motivation behind the model and other related literature. In Section 3 we analyze the projection onto the unit sphere of the walk that is obtained by averaging over the additive noise. Then, in Section 4 we leverage the properties of this chain to obtain our main result on the convergence rate and cutoff.

2. Some background

2.1. A statistical motivation. Our interest in the problem arose from a certain Bayesian scheme introduced by de Finetti [3] to estimate posterior measures arising from partially exchangeable sets of data. We refer to the last chapter of [13] for
more background. Given a connected network on $d$ vertices with no loops and with weights $c_{ij}$ together with $x^* \in \mathbb{R}^d$ and $g \in \mathbb{R}_+$, consider the quadratic form

$$Q(x) = \sum_{1 \leq i < j \leq d} c_{ij}(x_i - x_j)^2 + \sum_{1 \leq i \leq d} g_i(x_i - x_i^*)^2$$

$$= \sum_{i=1}^d \frac{1}{\sigma_i^2} \left[ e_i (x_i - \hat{x}_i)^2 + (1 - c_i) (x_i - x_i^*)^2 \right],$$

where we defined

$$e_i = \frac{\sum_{j \neq i} c_{ij}}{\sum_{j \neq i} c_{ij} + g_i}, \quad \sigma_i^2 = \frac{1}{\sum_{j \neq i} c_{ij} + g_i}.$$

and $\hat{x}_i$ is defined via the transition probabilities on the network (i.e., $p_{ij} = \frac{c_{ij}}{\sum_{j \neq i} c_{ij}}$).

In its original work [3], de Finetti was interested in understanding samples from a truncated Gaussian measure on the unit cube $[0, 1]^d$ with density

$$\pi(x) \propto e^{-\frac{Q(x)}{2}},$$

where $n$ is a large parameter. To overcome numerical problems arising from the truncation [2], one can utilize a Gibbs sampler to sample from this measure. Standard concentration inequalities (see [13]) show that the mixing time of the Gibbs sampler is only mildly affected by the constraint for $n$ large, as long as $e_i \in (0, 1)$ for all $1 \leq i \leq d$. Then, up to a translation and dilation, the problem boils down to the understanding of the auto-regressive model with the $Z_k$'s being standard normals, and $|X_0| = O(\sqrt{n})$.

As a corollary of Theorem [11] we obtain that the mixing time for the Gibbs sampler associated to $\pi$ is of order $\ln n$ as long as $e_i > 0$. The case $e_i = 1$ behaves rather differently and the mixing time becomes instead of order $n$ (see [7]).

2.2. Related work. There has been substantial work to understand the evolution of dynamics similar to the current setup. If we disregard the additive noise, we see that the starting point is closely related to the seminal result on random matrix products [6].

Proposition 2.1 (Furstenberg-Kesten theorem, [6]). Let $(C_k)_{k=1}^\infty$ be a strictly stationary ergodic series of $d \times d$ matrices such that $\mathbb{E} \log^+ \|C_1\| < \infty$. Then the following limit exists almost surely:

$$\lambda_1 = \lim_{k \to \infty} \frac{1}{k} \log \|C_k C_{k-1} \cdots C_1\| = \lim_{k \to \infty} \frac{1}{k} \mathbb{E} \log \|C_k C_{k-1} \cdots C_1\|.$$

In this expression $\lambda_1 < \infty$, but $\lambda_1 = -\infty$ may occur.

The cited result is general in terms of applicability, having minimal constraints on the matrix series. In the current setting, however, we want to understand the evolution at a finite horizon rather than in an asymptotic manner.

Substantial work on the discrepancy from the above limit rate has also been carried out. For an i.i.d. series of invertible matrices, $\frac{1}{k} \log \|C_k C_{k-1} \cdots C_1\| - k \lambda_1$ is asymptotically normal, as it was shown in [10] and refined in [1] where the optimal moment conditions were determined, namely $\mathbb{E} \log^2 \max(\|C_1\|, \|C_1^{-1}\|) < \infty$. Similar results are available when other structural requirements are made, in particular for allowable matrices. A non-negative matrix is allowable, if all rows and columns contain strictly positive elements. Together with additional assumptions, a central limit theorem is shown to hold for stationary ergodic random products of
such matrices (see [9]).

Observe that the set of matrices $A_i, 1 \leq i \leq d$ currently studied are neither invertible nor allowable, as the $i$th column of $A_i$ has all zero entries, which suggests the specialized challenge.

Moreover, our model (1.3) requires taking into account the additive term besides the linear map during the updates. One can consider a setup of even wider generality, by randomly iterating maps in a complete separable metric space $(S, \rho)$. That is, define a Markov chain using a collection of maps $\{f_\theta | \theta \in \Theta\}$ by

$$X_0 = x_0, \quad X_{k+1} = f_{\theta_{k+1}}(X_k),$$

with i.i.d. indices $\theta_k$ according to a distribution $\mu$. In this framework, stability can be ensured as follows.

**Proposition 2.2** ([5], Theorem 1.1.). In the above setup, assume for all $\theta \in \Theta$ that $f_\theta$ is Lipschitz with Lipschitz constant $K_\theta$. We further assume $\int_\Theta K_\theta \mu(d\theta) < \infty$, $\int_\Theta \log K_\theta \mu(d\theta) < 0$, and for some $x_0 \in S$, $\int_\Theta \rho(x_0, f_\theta(x_0)) \mu(d\theta) < \infty$.

Then the Markov chain in (2.1) has a unique stationary distribution, and exponential convergence occurs in the Prokhorov metric. Here, the rate is bounded away from 0 uniformly in $x_0$.

Moreover, the backward recursion $f_{\theta_1} \circ f_{\theta_2} \circ \ldots \circ f_{\theta_n}(x_0)$ converges almost surely.

This tool is powerful for its generality – several Markov chains can be cast in this language – and it highlights the contracting in average condition. While this can be relaxed in the affine case (see Theorem 2.1 in [5]) to include our case, it still does not capture exactly the rate as Proposition 2.1.

Therefore, our setup and claim fall outside the regime of the important works reviewed above.

3. A random walk on the sphere in the first quadrant

As already hinted at, the convergence rate of the chain $X_k$, as defined in (1.3), is essentially determined by the concentration properties of $Y_k$ defined via

$$Y_0 = \frac{X_0}{\|X_0\|}, \quad Y_k = \frac{A_k Y_{k-1}}{\|A_k Y_{k-1}\|}.$$  (3.1)

Notice that this is well defined as long as $X_0$ has positive coordinates, since the only effect of $A_k$ (defined in (1.2)) is to replace one coordinate with a damped weighted average of the others. It is worth noting that $Y_k$ is simply the chain obtained by averaging over the additive noise, and then normalized to lie on the unit sphere, i.e.,

$$Y_k = \frac{E_Z[X_k]}{\|E_Z[X_k]\|},$$  (3.2)

where $E_Z$ denotes the expectation with respect to $Z_1, \ldots, Z_k$.

Our first goal is to show a uniform bound on the ratio between coordinates in $Y_k$. We start by introducing some notation: for $y, y' \in \mathbb{R}^d$ write $y > y'$ or $y \geq y'$ if the same type of inequality holds coordinatewise. For $\delta \geq 0$, let $1_\delta$ be the vector with all components identically equal to $\delta$. Then, let

$$S_\delta := \{y \in \mathbb{R}^d | \|y\| = 1, y > 1_\delta\}.$$  

We write $S$ for $S_0$. Notice that our assumption on $X_0$ implies that $Y_k \in S$ for all $k$. When each $c_i$ in the definition (1.2) is equal to one, then convexity entails
that $\min \frac{Y_k}{\max Y_k}$ remain bounded away from zero. The first step is to generalize this to the case $\epsilon_i < 1$, namely by showing that regardless of the choices of the updates, $Y_k \in \mathcal{S}_I$ for some $\delta > 0$ that does not depend on $k$.

**Lemma 3.1.** Let $\epsilon := \min_{i \neq j} \epsilon_{i,j} > 0$, where $i \sim j$ denotes a pair for which $p_{ij} > 0$. Then, for all choices of the updates, one has the bound

$$\min \frac{Y_k}{\max Y_k} \geq \min \frac{Y_0}{\max Y_0} \epsilon^{d-1}. \quad (3.3)$$

**Proof.** We start by observing that, since the statement we aim to prove is scale-invariant, we can study the chain $Y_k$ where we neglect the normalization in (3.1). For convenience, we still denote it by $Y_k$.

Moreover, if for another chain $\tilde{Y}_k$ we have $Y_0 \geq \tilde{Y}_0$ and the same updates are used for both chains, then $Y_k \geq \tilde{Y}_k$ for all $k$. Together with the observation

$$1_{\min Y_0} \leq Y_0 \leq 1_{\max Y_0},$$

it suffices to prove the statement when $Y_0 = 1_1$, for which the right side of (3.3) is just $\epsilon^{d-1}$. We now proceed by induction on $k$ as follows.

For all $1 \leq i \leq d$ we have

$$1 = (Y_0)_i \geq (\epsilon_i \tilde{Y}_0)_i = \epsilon_i.$$

We will now prove that the inequality in the middle holds for $Y_1$ as well, regardless of the choice of the updated index $I$. In fact, there are three possibilities:

- If $i \not\sim I$, then our assumption leads to

$$(Y_1)_i = (Y_0)_i \geq (\epsilon_i \tilde{Y}_0)_i = (\epsilon_i \tilde{Y}_1)_i.$$

- If $i = I$, then

$$(Y_1)_I = (\epsilon_I \tilde{Y}_0)_I = (\epsilon_I \tilde{Y}_1)_I.$$

- If $i \sim I$, then

$$(Y_1)_i = (Y_0)_i \geq (\epsilon_i \tilde{Y}_0)_i = (\epsilon_i \tilde{Y}_1)_i + \epsilon_i p_{iI} \left(-\epsilon_i \tilde{Y}_0)_i + (Y_0)_I \right) \geq (\epsilon_i \tilde{Y}_1)_i.$$

Iterating the argument above $k$ times for the subsequent updates, we obtain that for all choices of the updates and for all coordinates $1 \leq i \leq d$,

$$(Y_k)_i \geq (\epsilon_i \tilde{Y}_k)_i.$$

For any choice of $j \sim i$, owing to the definition (1.1) and our choice of $\epsilon$ we can bound

$$(Y_k)_i \geq \epsilon(Y_k)_i.$$

Consider now connecting the extremal coordinates $\text{argmin} Y_k = i_1 \sim i_2 \cdots i_{t-1} \sim i_t = \text{argmax} Y_k$, using a shortest path. Then, iterating the inequality above we obtain

$$\min Y_k \geq \epsilon^{t-1} \max Y_k,$$

and we conclude since $t \leq d$. \hfill $\square$

**Remark 3.2.** Here is a simple geometric interpretation of the proof. Each of the $A_i$’s projects a point – in a non-orthogonal fashion – onto some hyperplane $H_i$. We exploit that the connected component of $\mathbb{R}^d \setminus \bigcup_{i=1}^t H_i$ containing $1_1$ is invariant under our dynamic.
Remark 3.3. The inequality above is sharp when the underlying network structure is a line path of length $d$.

Armed with this lemma, our next step is to show that the law of $Y_k$ converges to a unique measure, independently of the starting position $Y_0 \in \mathcal{S}$.

3.1. Weak contraction in the Hilbert metric. Consider the Hilbert metric $h$ on $\mathcal{S}$, given by

$$h(y, y') := \ln \left( \frac{\max \frac{\mu}{\nu}}{\min \frac{\mu}{\nu}} \right).$$

Consider also the corresponding Wasserstein metric induced on Borel probability measures on $\mathcal{S}$

$$W(\mu, \nu) = \inf_{Y \sim \mu, Y' \sim \nu} \mathbb{E} \left( h(Y, Y') \right),$$

where the infimum is taken over all couplings $Y \sim \mu, Y' \sim \nu$.

Let us highlight a few properties of these metrics: the space $\mathcal{S}_\delta$ is compact for all $\delta > 0$, but not for $\delta = 0$, when equipped with the metric $h$. Moreover, convergence in the metric $W$ on the space of Borel measures on $\mathcal{S}_\delta$, $\delta > 0$, is tantamount weak convergence, owing to the boundedness of the metric. In particular, Prokhorov’s theorem guarantees that the compactness property is inherited by the Wasserstein space.

This allows us to prove the following result.

Lemma 3.4. There exists a unique limit $\nu$ for the law of $Y_k$, which is independent of the choice of $Y_0 \in \mathcal{S}$.

Proof. Let $Y_0 \in \mathcal{S}$ be arbitrary and let $\delta$ be small enough so that $Y_k \in \mathcal{S}_\delta$ for all $k$, which we can ensure owing to Lemma 3.1. A compactness argument yields immediately the existence of a limiting measure $\nu$ up to subsequences.

In order to show uniqueness, regardless of the initial condition, it suffices to show that for any pair of measures $\nu \neq \nu'$ on $\mathcal{S}_\delta$ we have a weak contraction between $\nu$ and $\nu'$ after $d$ steps of the Markov chain, i.e.,

$$W(\nu, \nu') = W(\nu_d, \nu'_d). \quad (3.4)$$

Here, $\nu_d, \nu'_d$ denote the laws of $Y_d, Y'_d$ with $Y_0, Y'_0$ being distributed according to $\nu, \nu'$. Indeed, if both $\nu \neq \nu'$ were stationary measures, then we would obtain

$$W(\nu, \nu') = W(\nu_d, \nu'_d) < W(\nu, \nu'),$$

which is a contradiction, and thus $\nu = \nu'$.

Owing to the convexity of the Wasserstein metric, it suffices to show the bound \[3.4\] for $\nu$ and $\nu'$ being delta masses at some $Y_0$ and $Y'_0$, in which case the right side becomes $h(Y_0, Y'_0)$ for some $Y_0$ and $Y'_0$ in $\mathcal{S}_\delta$ for some $\delta$.

Consider now the coupling where the same coordinates are updated for both $Y_0$ and $Y'_0$. After one step, the ratio $(Y_1)_i/(Y'_1)_i$, either remains the same (if coordinate $i$ is not selected) or otherwise is equal to the ratio of a weighted average of all other coordinates. In both cases, we have

$$\min \frac{Y_0}{Y_0} \leq \frac{(Y_1)_i}{(Y'_1)_i} \leq \max \frac{Y_0}{Y_0}.$$
Iterating, we obtain for all choices of indices
\[ \min \frac{Y_0}{Y_0'} \leq \min \frac{Y_d}{Y_d'} \leq \max \frac{Y_d}{Y_d'} \leq \max \frac{Y_0}{Y_0'}, \]
which implies \( h(Y_d, Y_d') \leq h(Y_0, Y_0') \) for all choices of the updated indices.

Moreover, there exists a selection of indices \( I_1, \ldots, I_d \) for which the rightmost inequality is strict. To show this, let \( I \) be the set of indices where the maximum \( \max \frac{Y_0}{Y_0'} \) is achieved. Notice that \( |I| \leq d \) (owing to the assumption that \( Y_0 \) and \( Y_0' \) are distinct), and that at least one element \( I_d \) of \( I \) is connected to an element of \( I^c \) (since the network is connected). Therefore, if we start by selecting \( I \), and \( Y_1^{I_d}, Y_1^{I_d'} \), we obtain the conclusion.

Let us denote the event of a specific such index series occurring by \( A \), and the corresponding instance of the Markov chain after \( d \) steps by \( Y_d \cap A \), \( Y_d' \cap A \). Similarly, the event for any other index series is denoted by \( A^c \), and the corresponding conditional version of the Markov chain by \( Y_d \cap A^c \), \( Y_d' \cap A^c \). Our previous observations entail
\[ h(Y_d, Y_d') < h(Y_0, Y_0'), \quad h(Y_d, Y_d') \leq h(Y_0, Y_0'), \]
so that we obtain
\[ W(\nu_d, \nu_d') \leq \left( 1 - \frac{1}{d^2} \right) h(Y_d, Y_d') + \frac{1}{d^2} h(Y_d, Y_d') \]
\[ < h(Y_0, Y_0') \]
\[ = W(\nu, \nu') \]
as desired.

\[ \square \]

**Remark 3.5.** The choice of \( Y_0 \) with identical coordinates show that the unique stationary measure has support contained in \( S_{d-1} \), for \( \epsilon \) defined in Lemma 3.1.

### 3.2. Concentration inequalities.

Let \( Y_0 \) denote a random variable distributed according to \( \nu \), the unique stationary measure given by Lemma 3.4. As observed in Remark 3.5, we obtain that \( Y_0 \in S_{d-1} \) with probability one. In particular, this shows that
\[ \alpha := \mathbb{E}[\ln \|A_1 Y_0\|] \in (-\infty, 0), \quad (3.5) \]
where the expectation is taken over \( Y_0 \sim \nu \) and \( I \) uniformly distributed in \( \{1, \ldots, d\} \). Since \( \nu \) is stationary we deduce
\[ Y_1 = \frac{A_1 Y_0}{\|A_1 Y_0\|} = \frac{Y_0}{\|Y_0\|}. \]
and more generally \( Y_k = \frac{Y_0}{\|Y_0\|} \). If \( X_0 = \|X_0\| Y_0 \), combining the above with (3.2) we obtain
\[ \mathbb{E}[\ln \|\mathbb{E}_Z[X_k]\|] - \ln |X_0| = \sum_{j=1}^{k} \mathbb{E}[\ln |A_{I_j} Y_j|] \]
\[ = k\mathbb{E}[\ln \|A_{I_1} Y_0\|] \]
\[ = ka. \]

Armed with this, we can prove the following.
Lemma 3.6. Let $Y_0 \in S_\delta$ and $X_0 = nY_0$. Then, there exists a constant $\gamma > 0$, independent of $k$ and $n$, such that

$$\mathbb{P} \left( \left| \ln \left( \frac{\|E_{z_1, \ldots, z_k}[X_k]\|}{n} \right) - k\alpha \right| \geq t\sqrt{k} \right) \leq 2e^{-\gamma t^2}$$

for all $t > 0$ and $k$ large enough.

Proof. In what follows, the symbol $\lesssim$ denotes an inequality up to a constant independent of $k, n$ and the choice of indices. Construct $X'_k$ from $X_k$ by re-sampling the $s$th update for some $1 \leq s \leq k$. Then

$$\frac{\|E_{z_1, \ldots, z_k}[X_k]\|}{\|E_{z_1, \ldots, z_k}[X'_k]\|} = \frac{\|E_{z_{s+1}, \ldots, z_k}[A_{t_1}Y_0]\|}{\|E_{z_{s+1}, \ldots, z_k}[A_{t_1}Y_0]\|}$$

for some $\tilde{Y}_0 \in S_{\delta'}$, where $\delta'$ depends on $\delta$ and $\epsilon$ only owing to Lemma 3.1. Since for all indices $I$

$$1 \lesssim A_t \tilde{Y}_0 \lesssim 1,$$

we can use Lemma 3.1 applied to $1$ to deduce

$$1 \lesssim \frac{\|E_{z_1, \ldots, z_k}[X_k]\|}{\|E_{z_1, \ldots, z_k}[X'_k]\|} \lesssim 1$$

or, equivalently,

$$1 \lesssim \left| \ln \|E_{z_1, \ldots, z_k}[X_k]\| - \ln \|E_{z_1, \ldots, z_k}[X'_k]\| \right| \lesssim 1.$$

Using the bounded difference inequality \[12\], we obtain the claim where $k\alpha$ is replaced by $E \ln \left( \frac{\|E_{z_1}[X]\|}{\|X_0\|} \right)$. Therefore, it suffices to show that for $n = 1$ one has

$$\left| E \ln \|E_{z_1, \ldots, z_k}[X_k]\| - k\alpha \right| \lesssim 1,$$

for then the claim follows by possibly decreasing $\gamma$ and taking $k$ large enough. Using the definition of $\alpha$, we obtain

$$E \ln \|E_{z_1, \ldots, z_k}[X_k]\| - k\alpha = E \left( \ln \frac{\|E_{z_1, \ldots, z_k}[X_k]\|}{\|E_{z_1, \ldots, z_k}[X'_k]\|} \right).$$

where $X'_k$ is distributed according to the stationary distribution $\nu$, and we use the same updates on both $X_k$ and $X'_k$. Since $X'_k \in S_{\nu,1}$ with probability one (see Remark 3.5), the claim then follows applying once more Lemma 3.1 as before. \(\square\)

4. PROOF OF THE MAIN RESULT

As hinted at in the introduction, the Markov chain $X_k$ defined in \[1.3\] has a unique stationary distribution, namely the law of $\overline{X}$ given by \[1.4\]. In particular, for any choice of $X_0$ one has

$$X_k = A_{t_k} \cdots A_{t_1}X_0 = A_{t_k} \cdots A_{t_1}b_{t_1}(Z_1) + A_{t_k} \cdots A_{t_1}b_{t_2}(Z_2) + \ldots + b_{t_k}(Z_k) \overset{d}{=} X_{t_1} \cdots X_{t_{k-1}}(Z_{k-1}) + \ldots + b_{t_k}(Z_1),$$

where we used exchangeability of the sequences $I_1, \ldots, I_k$ and $Z_1, \ldots, Z_k$. This entails

$$X_k = A_{t_k} \cdots A_{t_1}X_0 \overset{d}{\rightarrow} \overline{X}. \quad (4.1)$$

We are now ready to prove our main result.
Proof of Theorem 1.1. We start proving the first claim, namely
\[
\lim_{\beta \to +\infty} \lim_{n \to +\infty} d_{\text{tv}}(\pi, \pi_k) = 1.
\]

By definition of total variation distance, it is enough to show that for all \( \epsilon > 0 \) there exists \( \beta \in \mathbb{R} \) and \( R > 0 \) such that \( \mathbb{P}(X \in B_R) \to 0 \) and \( \mathbb{P}(X_k = A_{I_k} \ldots A_{I_1}, X_0 \in B_R) \to 1 - \epsilon \) for all \( k = k(n, \beta) \) (given in (1.5)) with \( n \) large enough. Here, \( B_R \) denotes the ball centered at the origin with radius \( R \).

Fix \( \epsilon > 0 \), and pick \( R \) large enough so that for all \( k \) sufficiently large
\[
\mathbb{P}(X \in B_R) \geq 1 - \frac{\epsilon}{3}, \quad \mathbb{P}(X_k = A_{I_k} \ldots A_{I_1}, X_0 \in B_R) \geq 1 - \frac{\epsilon}{3}.
\]

This is possible owing to (4.1). A union bound leads to
\[
\mathbb{P}(X_k \in B_R) \leq \mathbb{P}(X_k = A_{I_k} \ldots A_{I_1}, X_0 \not\in B_R) + \mathbb{P}(A_{I_k} \ldots A_{I_1}, X_0 \in B_{2R}) - \frac{\epsilon}{3} + \mathbb{P}(A_{I_k} \ldots A_{I_1}, X_0 \in B_{2R}).
\]

Therefore, we have
\[
\mathbb{P}(X \in B_R) - \mathbb{P}(X_k \in B_R) \geq 1 - \frac{2\epsilon}{3} - \mathbb{P}(A_{I_k} \ldots A_{I_1}, X_0 \in B_{2R})
\]
so that we obtain the claim provided that
\[
\mathbb{P}(A_{I_k} \ldots A_{I_1}, X_0 \in B_{2R}) \leq \frac{\epsilon}{3}
\]
for \( k \) as in (1.5) with \( \|X_0\| = n \) and \( \beta \) small enough. Since \( Z \) has mean zero, we have
\[
\mathbb{E}_Z[X_k] = A_{I_k} \ldots A_{I_1}, X_0,
\]
and passing to logarithms we need to bound
\[
\mathbb{P}(\ln \|\mathbb{E}_Z[X_k]\| \leq \ln(2R))
\]
Thanks to Lemma 3.6 we know that for all \( t \geq 0 \) and all \( k \) sufficiently large
\[
\mathbb{P}(\ln \|\mathbb{E}_Z[X_k]\| \leq \ln n + k\alpha - t\sqrt{\kappa}) \leq 2e^{-\gamma t^2}.
\]
In order to conclude, take \( t \) large enough so that the right side is smaller than \( \frac{\epsilon}{3} \).
Then, for \( k \) as in (1.5) we have
\[
\ln n + k\alpha - t\sqrt{\kappa} = \sqrt{\ln n} \left( -\beta - \frac{t}{\sqrt{-\alpha}} + o(1) \right) \geq \ln(2R)
\]
for \( \beta \) negative and with a large enough absolute value. Therefore,
\[
\mathbb{P}(\ln \|\mathbb{E}_Z[X_k]\| \leq \ln(2R)) \leq \frac{\epsilon}{3}
\]
for all \( k = k(n, \beta) \) with \( n \) sufficiently large, as desired.

We now move to the second claim, namely
\[
\lim_{\beta \to +\infty} \lim_{n \to +\infty} d_{\text{tv}}(\pi, \pi_k) = 0.
\]
Consider an arbitrary \( X_0 \in \mathcal{S}_\delta \) for some \( \delta > 0 \) with \( \|X_0\| = n \), and let \( X'_0 \) be distributed according to the stationary distribution (notice that \( \|X'_0\| = O(1) \)). Owing to the coupling interpretation of total variation distance we need to show that, for all \( \epsilon > 0 \), one can construct a coupling between \( X_k \) and \( X'_k \) such that
\[
\mathbb{P}(X_k \neq X'_k) \leq \epsilon
\]
for \( k = k(n, \beta) \) as in (1.5) with \( \beta \) large enough and all \( n \) sufficiently large.
Let $T$ denote the first time that all coordinates have been selected at least once. On \{ $T \geq k$ \}, let the two chains $X_k$ and $X'_k$ run independently. Conditioned on \{ $T \leq k$ \}, let $1 \leq k_1 \leq k$ be the last time that coordinate $i$ is selected, $i \in \{1, \ldots, d\}$. Consider a coupling between $X_k$ and $X'_k$ with the same choice of coordinate updates, and using the same additive noise except at the times $k_i$. Without loss of generality, assume that $k_1 < k_2 < \ldots < k_d = k$. Then we can write

$$X_k - X'_k = A_{k_1} \ldots A_{k_d} (X_0 - X'_0) + G(Z_{k_1} - Z'_{k_1}, \ldots, Z_{k_d} - Z'_{k_d}),$$

where $G = G_{I_1, \ldots, I_k}$ is the linear map that sends $z \in \mathbb{R}^d$ to

$$G(z) = \sum_{i=1}^d A_{k_i} \ldots A_{k_{i+1}} b_{I_{k_i}}(z).$$

Notice that the last summand reduces to $(0, \ldots, 0, \sigma_d z_d)$, and in general the $i$th summand is a vector with the first $i - 1$ entries being zero owing to (1.2) and the assumption $k_1 < \ldots < k_d$. In particular, the matrix $G$ is lower triangular with $\sigma_s$ on the diagonal. In particular, $G$ is invertible and its inverse has a uniformly bounded norm (with respect to $k$ and the choice of the indices).

Moreover, for all choices of $r > 0$ we have

$$d_{TV}(\pi_k, \pi) \leq \mathbb{P}(X_k \neq X'_k)$$

$$\leq \mathbb{P}(T > k) + \mathbb{P}([A_{k_1} \ldots A_{k_d} (X_0 - X'_0)] \geq r)$$

$$+ \sup_{|s| \leq r, I_1, \ldots, I_k} \mathbb{P}(Z \neq Z' + G^{-1}(s)).$$

Here, $Z$ and $Z'$ are vectors in $\mathbb{R}^d$ with i.i.d. components distributed according to $\gamma$. The first terms is smaller than $\epsilon/3$ for $k$ large, owing to a classical coupon collector argument. As for the last term, we can couple $Z$ and $Z'$ optimally so that (here we identify $\gamma$ with its density)

$$\mathbb{P}(Z - Z' \neq G^{-1}(s)) \leq \frac{1}{2} \int_{\mathbb{R}^d} |\gamma(z_1) \ldots \gamma(z_d) - \gamma(z_1 + G^{-1}_1(s)) \ldots \gamma(z_d + G^{-1}_d(s))| dz$$

$$\leq \frac{d}{2} \sup_{i \in \{1, \ldots, d\}} \int_{\mathbb{R}} |\gamma(z) - \gamma(z + G^{-1}_i(s))| dz$$

$$\leq \epsilon/3$$

provided that $s \leq r$ with $r$ small enough, owing to the uniform control on the inverse of $G$ and to the continuity of the translation operator on integrable functions. As for the second term, a union bounds yields

$$\mathbb{P}([A_{k_1} \ldots A_{k_d} (X_0 - X'_0)] \geq r) \leq \mathbb{P}([A_{k_1} \ldots A_{k_d} X_0] \geq \frac{r}{2}) + \mathbb{P}([A_{k_1} \ldots A_{k_d} X_0] \geq \frac{r}{2}).$$

Since $\|X_0\| = O(1)$ and $|X_0| = n$ with $X_0 \in S_\delta$ for some $\delta > 0$, it is enough to show that the second term is smaller than $\epsilon/6$. On the other hand, using Lemma 3.6 and following the very same approach of the proof of the first claim, we obtain that the second term is smaller than $\epsilon/6$ for $k = k(n, \beta)$ as in (1.5) with $\beta$ large and $n$ sufficiently large.

All together, this implies the main result. \qed

Acknowledgment

We warmly thank Persi Diaconis for suggesting the problem being studied, and for his constant help and support. B. Gerencsér was supported by NRDI (National
Research, Development and Innovation Office) grant KKP 137490 and by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

**References**

[1] Y. Benoist and J.-F. Quint. Central limit theorem for linear groups. *The Annals of Probability*, 44(2), Mar. 2016.

[2] N. Chopin. Fast simulation of truncated Gaussian distributions. *Statistics and Computing*, 21(2):275–288, 2010.

[3] B. de Finetti. *Sur la condition d’” Équivalence partielle.”*. Actualités Scientifiques et Industrielles, 1938.

[4] P. Diaconis. The cutoff phenomenon in finite Markov chains. *Proceedings of the National Academy of Sciences*, 93(4):1659–1664, 1996.

[5] P. Diaconis and D. Freedman. Iterated random functions. *SIAM Review*, 41(1):45–76, Jan. 1999.

[6] H. Furstenberg and H. Kesten. Products of random matrices. *The Annals of Mathematical Statistics*, 31(2):457–469, June 1960.

[7] B. Gerencsér and A. Ottolini. Rates of convergence for Gibbs sampling in the analysis of almost exchangeable data. arXiv preprint arXiv:2010.15539, 2020.

[8] W. Gilks, S. Richardson, and D. S. (eds.). *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, 1995.

[9] H. Hennion. Limit theorems for products of positive random matrices. *The Annals of Probability*, 25(4), Oct. 1997.

[10] E. Le Page. Théorèmes limites pour les produits de matrices aléatoires. *Publications mathématiques et informatique de Rennes*, 1980.

[11] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2006.

[12] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics, 1989*, pages 148–188. Cambridge University Press, Aug. 1989.

[13] A. Ottolini. *Birthday Problems and Rates of Convergence for Gibbs Sampling*. Stanford University, 2021.

(B. Gerencsér) ALFRED RÉNYI INSTITUTE OF MATHEMATICS, RÉÁLTANODA UTCA 13-15, BUDAPEST 1053 (HU) AND EÖTVÖS TARLÓS UNIVERSITY, DEPARTMENT OF PROBABILITY AND STATISTICS, PÁZMÁNY PÉTER SÉTÁNY 1/C, BUDAPEST 1117 (HU)

*Email address: gerencser.balazs@renyi.hu*

(A. Ottolini) DEPARTMENT OF MATHEMATICS, UNIVERSITY OF WASHINGTON, SEATTLE WA 98102 (USA).

*Email address: ottolini@uw.edu*