Occupancy distributions in Markov chains via Doeblin's ergodicity coefficient

Stephen Chestnut[1] and Manuel E. Lladser[1,2†]

[1] *Department of Applied Mathematics, University of Colorado, Boulder, CO 80309-0526, The United States*
[2] *Corresponding author e-mail: manuel.lladser@colorado.edu*

We state and prove new properties about Doeblin's ergodicity coefficient for finite Markov chains. We show that this coefficient satisfies a sub-multiplicative type inequality (analogous to the Markov-Dobrushin's ergodicity coefficient), and provide a novel but elementary proof of Doeblin's characterization of weak-ergodicity for non-homogeneous chains. Using Doeblin's coefficient, we illustrate how to approximate a homogeneous but possibly non-stationary Markov chain of duration n by independent and short-lived realizations of an auxiliary chain of duration of order $\ln(n)$. This leads to approximations of occupancy distributions in homogeneous chains, which may be particularly useful when exact calculations via one-step methods or transfer matrices are impractical, and when asymptotic approximations may not be yet reliable. Our findings may find applications to pattern problems in Markovian and non-Markovian sequences that are treatable via embedding techniques.

Keywords: compound Poisson approximation, Doeblin, Markov chain embedding technique, motif, occupancy distribution, pattern

1 Introduction.

In what follows, S is a given finite set and $T \subset S$ a certain non-empty subset of states. For a fixed integer $n \geq 1$, consider a first-order homogeneous Markov chain $X = (X_t)_{0 \leq t \leq n}$ with initial distribution $\mu : S \to [0, 1]$ and probability transition matrix $p : S \times S \to [0, 1]$. We identify complex-valued functions defined over S and $S \times S$ as row vectors and matrices, respectively. In particular, the distribution of X_t is given by the vector μp^t.

Our object of interest is the *occupancy distribution of* T i.e. the distribution of the random variable:

$$T_n = \sum_{t=1}^{n} [\![X_t \in T]\!], \tag{1}$$

where $[\![\cdot]\!]$ denotes the indicator function of the event within (Iverson's bracket). Random variables of this sort are common in assessing the frequency statistics of patterns in random sequences, which typically

[†] This research has been partially supported by NSF grant DMS #0805950.

model text or genomic sequences. Although various probabilistic and analytic techniques have been used for this purpose, the *Markov chain embedding technique* is among the most versatile ones. This technique seems to have originated in the works of Gerber and Li [GL81], Biggins and Cannings [BC87], and Bender and Kochman [BK93]. It consists in embedding a random sequence into the state space of a suitable finite automaton that is informative of the pattern of interest, and it has been completely systematized for *regular patterns* i.e. patterns described by a regular expression, and *Markovian models* of random sequences [NSF02, Nic03]. In addition, the technique has also shown some promise for assessing regular patterns in *non-Markovian sequences* i.e. sequences with an arbitrary correlation structure [Lla08].

All the complexity associated with determining or approximating the distribution of T_n is due to the distributional dependence between the consecutive states visited by the chain. There are various ways–some more ad hoc and others more systematic–to pinpoint this distribution. For small values of n, exact calculations are possible via one-step methods [Dur99] or transfer matrices [FS09]. Furthermore, transfer matrices lead to Normal approximations for large values of n e.g. as shown in [NSF02] for frequency statistics of regular patterns under Markovian models. On the other hand, for stationary chains, Poisson [BHJ92] and compound Poisson approximations [Erh99] have been proposed when T is a *rare set* i.e. the stationary measure of T is small. A specialized instance of these approximations is the Pólya-Aeppli distribution which occurs as the limiting distribution of frequency statistics of rare words under stationary Markov models [RS07].

Our goal is to approximate the distribution of T_n, when X is possibly non-stationary, and in the intermediate regime where n is perhaps too large for exact calculations but too small to rely on asymptotic approximations. Our approach relies on a novel use of the so called *Doeblin's ergodicity coefficient* associated with p [Doe37], which is defined as:

$$\alpha(p) \stackrel{def}{=} \sum_j \min_i p(i,j). \tag{2}$$

The above goal is far from artificial! For instance, extensive research is being performed to understand the evolution of complex but relatively short RNA sequences from simpler but functional RNA sequences [KLYK08, KLW$^+$10]. In contexts like this, the pitfall of the Normal approximation of T_n is the slow rate of convergence of order $n^{-1/2}$. On the other hand, the stationary assumption of the aforementioned Poisson approximations is unrealistic in the context of the Markov chain embedding technique, even if the background model of a genomic sequence is Markovian and stationary. For example, for regular patterns, the initial distribution of the embedded process is concentrated in a few states (of an exponentially large state space) associated with the unique initial state of the deterministic finite automaton that recognizes the pattern of interest [Lla07].

1.1 Ergodicity coefficients and inhomogeneous Markov chains.

This section finishes the Introduction with a brief discussion about ergodicity coefficients and the historical developments surrounding the characterization of weak-ergodicity of non-homogenous Markov chains.

In what follows we denote the set of all probability transition matrices over the state space S as $\mathcal{P}$. The set of all stochastic matrices with identical rows is denoted $\mathcal{E}$; in particular, $\mathcal{E} \subset \mathcal{P}$. We refer to matrices in $\mathcal{E}$ as i.i.d. models because a homogeneous Markov chain with a probability transition matrix in this set is just a sequence of independent and identically distributed (i.i.d.) S-valued random variables.

Broadly speaking, an *ergodicity coefficient* is any continuous function $\gamma : \mathcal{P} \to [0,1]$. Such function is said *proper* when $\gamma(p) = 1$ if and only if $p \in \mathcal{E}$. Clearly, Doeblin's coefficient as defined in (2) is proper.

Other ergodicity coefficients found in the literature are:

$$\begin{aligned}
\gamma_1(p) &\stackrel{def}{=} \max_j \min_i p(i,j);\\
\gamma_2(p) &\stackrel{def}{=} \min_{i,j} \sum_s \min\{p(i,s), p(j,s)\} = 1 - \frac{1}{2} \cdot \max_{i,j} \sum_s |p(i,s) - p(j,s)|;\\
\gamma_3(p) &\stackrel{def}{=} 1 - \max_s \max_{i,j} |p(i,s) - p(j,s)|.
\end{aligned}$$

Only the last two of these are proper. γ_2 is called the *Markov-Dobrushin's* ergodicity coefficient [Mar06, Dob56a, Dob56b].

Ergodicity coefficients have been proposed for a range of purposes such as to analyze the contractive property of a stochastic matrix [Mar06] and bound its non-Perron-Frobenius eigenvalues [Sen93]. However, they have mostly been used to analyze the asymptotic behavior of non-homogenous Markov chains [Doe37, Dob56a, Sen73b]. This entails understanding the asymptotic behavior of products of the form $\prod_{k=m}^{n} p_k$, with $m \le n$, for a given sequence $(p_k)_{k\ge 0} \subset \mathcal{P}$. Such a sequence is said to be *weakly ergodic* if for all $m \ge 0$ and $i,j,s \in S$ the following applies:

$$\lim_{n\to\infty} \left| \Big(\prod_{k=m}^{n} p_k\Big)(i,s) - \Big(\prod_{k=m}^{n} p_k\Big)(j,s) \right| = 0. \tag{3}$$

(This condition does not imply that $\prod_{k=m}^{n} p_k$ converges to a limit as n tends to infinity. As a counter-example, just consider the case with $p_k = E_0$ for k even, $p_k = E_1$ for k odd, for given but different matrices $E_0, E_1 \in \mathcal{E}$.)

The following condition, known as *Markov's theorem* [Sen73a], is sufficient for weak ergodicity:

$$\sum_{k=0}^{\infty} \gamma_1(p_k) = +\infty. \tag{4}$$

This condition is only sufficient in great part because γ_1 is not proper [Sen73b].

In more probabilistic terms, consider a first-order Markov chain $Y = (Y_k)_{k\ge 0}$ with state space S such that $\mathbb{P}[Y_k = s \mid Y_{k-1}, \ldots, Y_0] = p_k(Y_{k-1}, s)$, for each $k \ge 1$. The sequence $(p_k)_{k\ge 0}$ is weakly ergodic if and only if any two independent realizations of Y meet infinitely often on a same state with probability one. This characterization is due to Doeblin and appeared without proof in the report [Doe37]. The way in which Doeblin proved this result is matter of speculation and it was lost with his death in World War II (see [Sen73b] for the historical developments). Furthermore, Doeblin's report remained unnoticed for almost two decades. During this period, the following condition was proved to be both necessary and sufficient for weak ergodicity [Haj58]:

$$\left\{ \begin{array}{l} \text{there exists a strictly increasing sequence of positive} \\ \text{integers } (n_k)_{k\ge 0} \text{ such that: } \sum_{k=0}^{\infty} \gamma_2\Big(\prod_{i=n_k}^{n_{k+1}-1} p_i\Big) = +\infty. \end{array} \right. \tag{5}$$

In contrast, Doeblin's characterization of weak ergodicity is the following [Doe37]:

$$\begin{cases} \text{there exists a strictly increasing sequence of positive} \\ \text{integers } (n_k)_{k\geq 0} \text{ such that: } \sum_{k=0}^{\infty} \alpha\Big(\prod_{i=n_k}^{n_{k+1}-1} p_i\Big) = +\infty. \end{cases} \tag{6}$$

Since $\gamma_1(p) \leq \alpha(p) \leq \gamma_2(p)$, for each $p \in \mathcal{P}$, the sufficient condition in (4) is a special instance of the conditions in (5) and (6). Though nobody knows how Doeblin proved that conditions (3) and (6) are equivalent, Seneta ventured in [Sen73b] a possible proof. His proofs relies on the following two facts, valid any sequence $(p_k)_{k\geq 0} \subset \mathcal{P}$:

(a) $\Big(1 - \gamma_3\big(\prod_{k=0}^{n} p_k\big)\Big) \leq \prod_{k=0}^{n} \big(1 - \gamma_2(p_k)\big)$, for all $n \geq 0$;

(b) if $\sum_{k=0}^{\infty} \alpha\big(p_k\big) = +\infty$ then $\sum_{k=0}^{\infty} \gamma_1\big(p_k\big) = +\infty$.

Paper overview and organization. Our paper is mostly about Doeblin's ergodicity coefficient, which we encountered—by accident—while aiming at accurate but low-to-moderate complexity approximations of occupancy distributions in homogenous Markov chains. Here we mostly state and prove new properties about Doeblin's coefficient which we would have never explored otherwise. The more detailed implications of these properties to approximate occupancy distributions will be part of a follow up publication based on the M.S. thesis [Che10].

In §2 we demonstrate new properties about Doeblin's coefficient, which we use in §2.1 to provide a new and more elementary proof of Doeblin's characterization of weak-ergodicity of non-homogenous Markov chains. In §3, we relate Doeblin's coefficient to a decomposition of a chain into several independent realizations of an auxiliary chain. This decomposition has been used in the Markov chain Monte Carlo literature to sample from the stationary distribution. In §3.1, we illustrate how this decomposition allows one to approximate the distribution of a particular X_n with $O(\ln(n))$ rather than n matrix multiplications. In §3.2, we go a step further and illustrate how the decomposition allows us to parse (with high probability) the trajectory of a Markov chain of duration n into short-lived realizations of an auxiliary chain of durations of order $\ln(n)$. We exploit this feature to propose new approximations for occupancy distributions based on Doeblin's coefficient, which we compare against Normal and Poisson approximations in a numerical example.

2 A candidate for Doeblin's missing proof.

Recall that Doeblin's ergodicity coefficient associated with a $p \in \mathcal{P}$ is the quantity:

$$\alpha(p) = \sum_j \min_i p(i,j).$$

Because $\alpha(\cdot)$ is a proper ergodicity coefficient, $\alpha(p)$ is close to 1 when p is in the proximity of some i.i.d. model. However, since the set of i.i.d. models is closed, there should be several i.i.d. models close to p. The following result identifies an affine space of i.i.d. models that are in the proximity of p.

Theorem 2.1 *For each $p \in \mathcal{P}$, the following applies:*

(a) If $0 \leq \alpha \leq \alpha(p)$ then there is $E \in \mathcal{E}$ and $M \in \mathcal{P}$ such that $p = \alpha \cdot E + (1-\alpha) \cdot M$.

(b) $\alpha(p) = \sup \left\{ \alpha \in [0,1] \,\middle|\, (\exists E \in \mathcal{E})(\exists M \in \mathcal{P})\colon p = \alpha{\cdot}E + (1-\alpha){\cdot}M \right\}$.

(c) Assume that $E \in \mathcal{E}$ and $M \in \mathcal{P}$ are such that $p = \alpha(p) \cdot E + \big(1 - \alpha(p)\big) \cdot M$.
If $\alpha(p) < 1$ then $\alpha(M) = 0$ i.e. M has a zero in each column.
If $\alpha(p) > 0$ then $E(i,j) = \frac{1}{\alpha(p)} \cdot \min\limits_{s} p(s,j)$.

Proof: Define $\beta = \alpha(p)$. We first show part (a) in the theorem, for which we may assume without loss of generality that $\beta > 0$. In this case, all reduces to prove that there is $E \in \mathcal{E}$ and $M \in \mathcal{P}$ such that

$$p = \beta \cdot E + (1-\beta) \cdot M. \tag{7}$$

Indeed, if $0 \leq \alpha \leq \beta$ then the above implies that $p = \alpha E + (\beta - \alpha)E + (1-\beta)M = \alpha E + (1-\alpha)Q$, for some $Q \in \mathcal{P}$, because the matrix $(\beta - \alpha)E + (1-\beta)M$ has nonnegative entries and the sum of the entries in each of its rows is $(1-\alpha)$. To prove the above identity, consider the matrix $E \in \mathcal{E}$ with entries $E(i,j) = \min_s p(s,j)/\beta$. Since $\beta E(i,j) \leq p(i,j)$, the matrix $(p - \beta E)$ has nonnegative entries and row sums equal to $(1-\beta)$. In particular, if $\beta = 1$ then $p = E$ and the above identity holds with any choice of M. Otherwise, it suffices to select $M = (p - \beta E)/(1-\beta)$. This shows (7) and completes the proof of part (a).

To show part (b), notice that if $0 \leq \alpha \leq 1$ is such that $p = \alpha E + (1-\alpha)M$, with $E \in \mathcal{E}$ and $M \in \mathcal{M}$, then $\beta = \alpha(p) \geq \alpha \cdot \alpha(E) = \alpha$. Part (b) is now a direct consequence of part (a).

Finally we show part (c). Thus assume that $E \in \mathcal{E}$ and $M \in \mathcal{P}$ are such that $p = \beta E + (1-\beta)M$. If $\beta = 1$ then $p = E$ and the identity $E(i,j) = \min_s p(s,j)/\beta$ is trivial. On the other hand, if $\beta = 0$ then p must have a zero in each column and $M = p$. Without loss of generality we may therefore assume that $0 < \beta < 1$. We first show that $\alpha(M) = 0$. Set $\alpha' = \alpha(M)$. Due to part (a), there exists $E' \in \mathcal{E}$ and $M' \in \mathcal{P}$ such that $p = \beta E + (1-\beta)\alpha' E' + (1-\beta)(1-\alpha')M'$. Hence $\beta = \alpha(p) \geq (\beta + (1-\beta)\alpha')$ and as a result $\alpha' = 0$. To complete the proof of the theorem, fix j and notice that $\beta E(i,j) = p(i,j) - (1-\beta)M(i,j)$. In particular, since M has a zero in each column, there is s such that $\beta E(s,j) = p(s,j)$. Finally, since $\beta E(i,j) \leq p(i,j)$, we conclude that $\beta E(i,j) = \min_s p(s,j)$. This completes the proof. $\square$

Due to part (a) in the previous theorem, for all $p \in \mathcal{P}$, there is $E \in \mathcal{E}$ and $M \in \mathcal{P}$ such that:

$$(p - E) = (1 - \alpha(p)) \cdot (M - E).$$

As a result, the smaller $(1 - \alpha(p))$, the closer p is to an i.i.d. model. According to the following corollary, $\big(1 - \alpha(\cdot)\big)$ is a sub-multiplicative function; in particular, when one multiplies two or more stochastic matrices, one can only get "closer" to the set of i.i.d. models. This is the key ingredient for our proof of Doeblin's characterization of weak ergodicity.

To the best of our knowledge, the sub-multiplicative property has been shown in the literature only for the Markov-Dobrushin ergodicity coefficient [Dob56a, Dob56b, Ios72, Paz70, Gri75].

Corollary 2.2 $\big(1 - \alpha(pq)\big) \leq \big(1 - \alpha(p)\big) \cdot \big(1 - \alpha(q)\big)$*, for all $p, q \in \mathcal{P}$.*

Proof: Define $\alpha_1 = \alpha(p)$ and $\alpha_2 = \alpha(q)$. Due to part (a) in Theorem 2.1, there are matrices $E_1, E_2 \in \mathcal{E}$ and $M_1, M_2 \in \mathcal{P}$ such that $p = \alpha_1 E_1 + (1-\alpha_1)M_1$ and $q = \alpha_2 E_2 + (1-\alpha_2)M_2$. In particular, since $pE_2 = E_2$, we see that $pq = \alpha_2 E_2 + \alpha_1(1-\alpha_2)E_1M_2 + (1-\alpha_1)(1-\alpha_2)M_1M_2$. But notice that $E_1M_2 \in \mathcal{E}$. Consequently, the rows of the matrix $\alpha_2 E_2 + \alpha_1(1-\alpha_2)E_1M_2$ are identical, with common row sum $(\alpha_1+\alpha_2-\alpha_1\alpha_2)$. As a result, there is $E_3 \in \mathcal{E}$ such that

$$\begin{aligned} pq &= (\alpha_1+\alpha_2-\alpha_1\alpha_2)\cdot E_3 + (1-\alpha_1)(1-\alpha_2)\cdot M_1M_2; \\ &= (1-(1-\alpha_1)(1-\alpha_2))\cdot E_3 + (1-\alpha_1)(1-\alpha_2)\cdot M_1M_2. \end{aligned}$$

Finally, due to part (b) in Theorem 2.1, it follows from the above that

$$\alpha(pq) \geq 1-(1-\alpha_1)(1-\alpha_2),$$

which proves the corollary. □

2.1 A first principles proof of Doeblin's characterization of weak ergodicity.

As we mentioned earlier, Doeblin's proof that (3) and (6) are equivalent is matter of speculation. Though it is possible to prove this equivalence using Theorem 1 in [Sen73b] and Corollary 2.2 in the previous section, we venture an alternative and more elementary proof of this fact.

Fix an integer $m \geq 0$ and let α_n denote the Doeblin's ergodicity coefficient of $\prod_{k=m}^n p_k$. Due to parts (a) and (c) in Theorem 2.1, there are matrices $E_n \in \mathcal{E}$ and $M_n \in \mathcal{P}$ such that:

$$\prod_{k=m}^{n} p_k = \alpha_n \cdot E_n + (1-\alpha_n)\cdot M_n, \text{ with } \alpha(M_n) = 0.$$

In particular, for all $i, j, s \in S$ the following holds:

$$\Big(\prod_{k=m}^{n} p_k\Big)(i,s) - \Big(\prod_{k=m}^{n} p_k\Big)(j,s) = (1-\alpha_n)\cdot\big(M_n(i,s) - M_n(j,s)\big). \tag{8}$$

Assume first that condition (6) holds. Consider the sets of non-negative integers:

$$\begin{aligned} I_n &= \{k \mid \exists\, j \text{ such that } m \leq n_j \leq k \leq n_{j+1} \leq n\}; \\ J_n &= \{j \mid \{n_j, \ldots, n_{j+1}\} \subset I_n\}. \end{aligned}$$

Notice that J_n is an interval of integers. Furthermore, there are stochastic matrices L and R_n such that $\prod_{k=m}^n p_k = L \cdot \big(\prod_{k\in I_n} p_k\big) \cdot R_n$. In particular, due to Corollary 2.2, we find that

$$(1-\alpha_n) \leq \Big(1-\alpha\big(\prod_{k\in I_n} p_k\big)\Big) \leq \prod_{j\in J_n}\Big(1-\alpha\big(\prod_{k=n_j}^{n_{j+1}-1} p_k\big)\Big).$$

Since $(1-x) \leq \exp(-x)$, for $0 \leq x \leq 1$, the condition in (6) implies that $\lim\limits_{n\to\infty} \alpha_n = 1$. Back in (8), since each M_n is a stochastic matrix, we conclude that

$$\lim_{n\to\infty} \Big(\prod_{k=m}^{n} p_k\Big)(i,s) - \Big(\prod_{k=m}^{n} p_k\Big)(j,s) = 0.$$

This shows that condition (3) is also satisfied i.e. $(p_k)_{k\geq 0}$ is weakly ergodic.

Conversely, assume that condition (3) holds. To show that condition (6) also applies, we first prove that $(\alpha_n)_{n\geq 0}$ has a subsequence that converges to one. We show this by contradiction. In particular, due to the identity in (8), it applies that

$$\lim_{n\to\infty} \big(M_n(i,s) - M_n(j,s)\big) = 0,$$

for all $i,j,s \in S$. Fix $s_1 \in S$. Since each M_n has at least one zero in the column associated with s_1 then there is $j_1 \in S$ and a subsequence $(n'_k)_{k\geq 0}$ such that $M_{n'_k}(j_1,s_1) = 0$ for all $k \geq 0$. Therefore, $M_{n'_k}(i,s_1) \to 0$ as $k \to \infty$, for all $i \in S$. Now fix $s_2 \in S \setminus \{s_1\}$. Since each $M_{n'_k}$ has at least one zero in the column associated with s_2 then there is a subsequence $(n''_k)_{k\geq 0}$ of $(n'_k)_{k\geq 0}$ such that $M_{n''_k}(i,s_1) \to 0$ and $M_{n''_k}(i,s_2) \to 0$ as $k \to \infty$, for all $i \in S$. Since S is finite, a straightforward inductive argument shows that there is a subsequence $(n_k)_{k\geq 0}$ such that

$$\lim_{k\to\infty} M_{n_k}(i,s) = 0,$$

for all $i,s \in S$. However, the above is not possible because each M_{n_k} is a stochastic matrix. As a result, $(\alpha_n)_{n\geq 0}$ must have a subsequence that converges to one.

The previous argument shows that if $(p_k)_{k\geq 0}$ is weakly ergodic then for all $m \geq 0$ there is $n \geq m$ such that e.g. $\alpha(\prod_{k=m}^{n} p_k) \geq 1/2$. From this, condition (6) is immediate and we have proved that conditions (3) and (6) are equivalent.

3 Occupancy distributions of homogeneous chains.

In this section, we retake our initial motivation of approximating occupancy distributions of homogeneous Markov chains. As a starting point, suppose that:

$$p = \alpha \cdot E + (1-\alpha) \cdot M, \tag{9}$$

for certain $0 \leq \alpha \leq 1$, $E \in \mathcal{E}$ and $M \in \mathcal{P}$, and denote as $\mathbf{e}$ any of the rows of E (see Theorem 2.1). The following stochastic equivalent of $X = (X_i)_{i\geq 0}$ has been exploited in the Markov chain Monte Carlo literature to sample—perfectly—from the stationary distribution, without having to compute it beforehand [MG98, Møl99, CT01].

Lemma 3.1 *Assume that condition (9) is satisfied. Imagine a coin that shows E with probability α and M with probability $(1-\alpha)$ when tossed. The stochastic sequence $Y = (Y_i)_{i\geq 0}$ defined as follows:*

(i) Y_0 has distribution μ, and

(ii) for each $i \geq 0$, the distribution of Y_{i+1} conditioned on $(Y_0,\ldots,Y_i)$ is given by the following procedure: toss the coin, and if the E-side comes up then draw Y_{i+1} using the distribution $\mathbf{e}(\cdot)$, else draw Y_{i+1} using the distribution $M(Y_i,\cdot)$,

has the same distribution as X.

Proof: Due to the definition of the Y process, for each $i \geq 0$ and $s_0, \ldots, s_{i+1} \in S$, the following applies:

$$\begin{aligned}\mathbb{P}(Y_{i+1} = s_{i+1} \mid Y_0 = s_0, \ldots, Y_i = s_i) &= \alpha \cdot e(s_{i+1}) + (1-\alpha) \cdot M(s_i, s_{i+1}), \\ &= \alpha \cdot E(s_i, s_{i+1}) + (1-\alpha) \cdot M(s_i, s_{i+1}) = p(s_i, s_{i+1}).\end{aligned}$$

In particular, Y is a first-order homogeneous Markov chain with initial distribution μ and probability transition matrix p. Hence X and Y have the same distribution. $\square$

Rather than simulating perfectly from the stationary distribution, we illustrate how to use Lemma 3.1 to approximate occupancy distributions in homogeneous Markov chains. The key idea is to use Doeblin's ergodicity coefficient to break—at random times—the dependence of the chain. As an intermediate step, we consider a simpler problem: how to approximate the distribution of a particular X_n. Of course, if the chain is aperiodic and irreducible then, for large n, the distribution of X_n will be well-approximated by the stationary distribution. However, as pointed out in §1, our goal is to provide approximations in the intermediate regime where asymptotic approximations may not yet be reliable.

For what follows, recall that the *total variation distance* between two probability distributions $u(\cdot)$ and $v(\cdot)$ supported over $\mathbb{N} = \{0, 1, \ldots\}$ is defined as:

$$\|u - v\| \stackrel{def}{=} \sup_{A \subset \mathbb{N}} |u(A) - v(A)| = \frac{1}{2} \sum_{i \in \mathbb{N}} |u(\{i\}) - v(\{i\})|.$$

Accordingly, the total variation distance between two $\mathbb{N}$-valued random variables U and V is defined as the total variation distance of their distributions and it is denoted $\|U - V\|$.

3.1 A simpler problem.

It is well-known that, in finite state spaces, if p is irreducible and aperiodic then there is a unique stationary distribution i.e. a unique probability distribution π such that $\pi p = \pi$. In this case, there are constants $c_0, c_1 > 0$ which depend on p but not on μ, such that:

$$\|X_n - \pi\| \leq c_0 e^{-c_1 \cdot n}, \text{ for all } n \geq 0. \tag{10}$$

The proof of this fact can be found in most books on Markov chain theory, with approaches ranging from the well-known *Perron-Frobenius Theorem* [Per07, Fro12] to Operators and Ergodic theory [Lin71]. Various proofs are based on the *coupling method*, generally attributed to Doeblin, who proved the above in [Doe38]. The coupling technique has been since then refined to address Markov chains with countable state spaces [Pit74, Gri75, Tho90] and a purely probabilistic proof of the Renewal theorem [Lin77].

Using Lemma 3.1, we may alternatively explain (10) as follows. If $\alpha(p) > 0$ then there is a distribution $\mathbf{e}$ such that, regardless of the state where the chain is located, the next state will be picked up from this distribution with probability $\alpha(p)$. Each time this distribution is used, any information about the states previously visited by the chain is lost. This distribution acts therefore as a "memory-breaker". When n is large, and even if $\alpha(p)$ is small, it is unlikely that no memory-breaker occurred between X_0 and X_n. Since all transitions after the last memory-breaker where controlled by M, the distribution of X_n should be well-approximated by a mixture of distributions of the form $\mathbf{e}M^t$. This intuition is made precise on the following result. Due to part (b) in Theorem 2.1, notice that the optimal choice for α is $\alpha(p)$.

Corollary 3.2 *[Che10] Assume that condition (9) is satisfied. Let $m \geq 0$ and consider S-valued random variables $Z_0, \ldots, Z_m$ such that Z_t has distribution $\mathbf{e}M^t$. If I is a random index independent of $(Z_1, \ldots, Z_m)$ such that $\mathbb{P}[I = t] = \alpha(1-\alpha)^t/(1-(1-\alpha)^{m+1})$, for $0 \leq t \leq m$; in particular, $\mathbb{P}[I \in \{0, \ldots, m\}] = 1$, then*

$$\|X_n - Z_I\| \leq (1-\alpha)^{m+1}, \textit{ for all } n > m.$$

To fix ideas, consider the probability transition matrix:

$$p = \begin{bmatrix} \frac{3}{10} & 0 & \frac{7}{10} \\ 0 & \frac{9}{10} & \frac{1}{10} \\ \frac{4}{5} & \frac{1}{5} & 0 \end{bmatrix}. \tag{11}$$

Since $\alpha(p) = 0$, the inequalities of the corollary are trivial. However, observe that:

$$p^2 = \begin{bmatrix} \frac{13}{20} & \frac{7}{50} & \frac{21}{100} \\ \frac{2}{25} & \frac{83}{100} & \frac{9}{100} \\ \frac{6}{25} & \frac{9}{50} & \frac{29}{50} \end{bmatrix} = \frac{31}{100} \cdot E_2 + \frac{69}{100} \cdot M_2,$$

with

$$E_2 := \begin{bmatrix} \frac{8}{31} & \frac{14}{31} & \frac{9}{31} \\ \frac{8}{31} & \frac{14}{31} & \frac{9}{31} \\ \frac{8}{31} & \frac{14}{31} & \frac{9}{31} \end{bmatrix} \text{ and } M_2 := \begin{bmatrix} \frac{19}{23} & 0 & \frac{4}{23} \\ 0 & 1 & 0 \\ \frac{16}{69} & \frac{4}{69} & \frac{49}{69} \end{bmatrix}.$$

Notice that $\alpha_2 := \alpha(p^2) = 31/100$; in particular, the above decomposition of p^2 is a direct consequence of part (c) in Theorem 2.1. Define $\mathbf{e}_2$ as the first row of E_2. Imagine you would like to approximate the distribution of some X_n, with as few matrix multiplications as possible, and within a 5% accuracy in total variation distance. Define $\epsilon := 0.05$. Due to the Corollary 3.2, this is possible for any even number $n \geq 18$, by considering a mixture of the distributions $\mathbf{e}_2\, M_2^t$, with $t = 0, \ldots, 8$. On the other hand, because Markovian kernels are contractive in total variation distance [Mar06], this is also possible for any odd number $n \geq 18$ by considering a mixture of the distributions $\mathbf{e}_2\, M_2^t\, p$, with $t = 0, \ldots, 8$. Either mixture can be computed in at most 10 matrix multiplications, however, as it follows from Table 1, this number can be optimized to a minimum of 7 matrix multiplications by considering larger powers of p.

According to the data displayed in Table 1, it is possible to approximate within ϵ-units the distribution of each X_n, with $n \geq 16$, using a mixture of three distributions associated with Doeblin's ergodicity coefficient of p^4. This mixture is given by:

$$\sum_{t=0}^{3} \frac{(1-\alpha_4)^t - (1-\alpha_4)^{t+1}}{1-(1-\alpha_4)^4} \cdot \mathbf{e}_4 \cdot M_4^t \cdot p^{n (\text{mod } 4)}, \tag{12}$$

which can be computed using no more than 7 matrix multiplications. In retrospect, this is far from obvious. For instance, one finds that:

$$p^7 = \begin{bmatrix} 0.3444507000 & 0.3440640000 & 0.3114853000 \\ 0.1966080000 & 0.6114381000 & 0.1919539000 \\ 0.3559832000 & 0.3839078000 & 0.2601090000 \end{bmatrix}.$$

Since $\max_{i,j} \|p^7(i,\cdot) - p^7(j,\cdot)\| \geq 0.26$, the chain is still far from the stationary distribution even after 7 transitions. Indeed, $\max_i \|p^7(i,\cdot) - \pi(\cdot)\| \geq 0.13$, exceeding the total variation distance between any X_n, with $n \geq 16$, and the distribution in (12).

k	α_k	m_k	n_k	c_k
1	0	∞	∞	∞
2	0.31	8	18	10
3	0.403	5	18	8
4	0.5287	3	16	7
5	0.60127	3	20	8
6	0.679891	2	18	8
7	0.7326259	2	21	9

Tab. 1: Parameters associated with powers of the probability transition matrix in (11). Here $\alpha_k = \alpha(p^k)$ and $p^k = \alpha_k E_k + (1-\alpha_k)M_k$, with $E_k \in \mathcal{E}$ and $M_k \in \mathcal{P}$. In addition, $m_k := \lceil \ln(\epsilon)/\ln(1-\alpha_k) - 1 \rceil$, with $\epsilon := 0.05$, and $n_k := k(m_k+1)$. Due to Corollary 3.2, if $\mathbf{e}_k$ denotes any of the rows of E_k then for each $n \geq n_k$ there exists a mixture of the distributions $\mathbf{e}_k\, M_k^t\, p^{n \pmod k}$; $t = 0, \ldots, m_k$, which approximates the distribution of X_n within ϵ-units in total variation distance. This mixture can be computed with at most $c_k := (k + m_k)$ matrix multiplications.

3.2 Approximation of occupancy distributions.

Assume that condition (9) is satisfied. Following the notation of Lemma 3.1, the occupancy distribution of a set $T \subset S$ is the distribution of the random variable:

$$T_n = \sum_{t=1}^{n} [\![Y_t \in T]\!].$$

The moment generating function (m.g.f.) of T_n is given by $\mu \cdot \{p(z)\}^n \cdot \mathbf{1}$, where $p(z)$ is the matrix with polynomial entries given by $p(z)(i,j) = p(i,j) \cdot z^{[\![j \in T]\!]}$, and $\mathbf{1}$ is a column-vector of ones. $p(z)$ is called a transfer matrix [FS09], and the computation of the exact distribution of T_n is expensive unless n is relatively small. In what follows, we extend the argument of the previous section to approximate this distribution.

Notice that the random variables $[\![Y_i \in T]\!]$ and $[\![Y_j \in T]\!]$, with $i < j$, are independent when at least one of the random variables $Y_{i+1}, \ldots, Y_j$ is drawn from the the memory-breaker distribution $\mathbf{e}$. In particular, the times at which the E-side of the coin appears cut the trajectory $Y_0, \ldots, Y_n$ into independent "pieces". The number of such pieces is random, and consecutive transitions in each piece are governed by the matrix M. Furthermore, the initial distribution of each piece is $\mathbf{e}$, except for the first piece which has initial distribution μ. The expected number of memory-breakers between the first and last transition is αn; and the average separation between consecutive memory-breakers is $1/\alpha$, regardless of n. As a result, a mixture of $\mathbf{e}(z) \cdot \{M(z)\}^m \cdot \mathbf{1}$, with m an integer in a neighborhood of $1/\alpha$, should lead to a decent approximation of the m.g.f. of the occupancy distribution of T in each piece other than the first. For the first piece, the m.g.f.'s to consider are of the form $\mu \cdot \{M(z)\}^m \cdot \mathbf{1}$. Since the behavior of the Markov chain is independent from one piece to another, an approximation for the m.g.f. of T_n should follow. More importantly for computations, a power of order $o(n)$ of the transfer matrix $M(z)$ should suffice for a decent approximation of the distribution of T_n. The weakest point of this heuristic is the probable occurrence of longer than expected pieces at already intermediate values of n. This motivates us to look at the random variable L_n defined as the length of the longest piece. (In probabilistic terms, L_n is the length of the largest run of M's in n-tosses of the coin from Lemma 3.1.) The asymptotic distribution of this random variable is well understood, both via combinatorial and probabilistic methods [Fel68, FS09, AGG90]. Since the distribution of L_n concentrates around $-\ln(\alpha n)/\ln(1-\alpha)$ as n increases, selecting $m = -c\ln(\alpha n)/\ln(1-\alpha)$, for $c > 1$, gives $\mathbb{P}[L_n \leq m] = 1 + \mathcal{O}(n^{1-c})$. An explicit upper-bound for the error in total variation distance follows now from the next result. We notice that the m.g.f. of the random variable W_I in the corollary can be computed explicitly via a symbolic specification [FS09].

Corollary 3.3 *[Che10] Assume that condition (9) is satisfied. Fix $m \geq 0$ and define $\ell_n = \mathbb{P}[L_n \leq m]$. Let $I = (I_0, \dots, I_K)$ be a random composition of n such that $\mathbb{P}[I = (i_0, i_1, \dots, i_k)] = \alpha^k(1-\alpha)^{n-k}/\ell_n$, for all $k \geq 0$, $0 \leq i_0 \leq m$, $1 \leq i_l \leq (m+1)$, for $l \geq 1$, and such that $\sum_{l=0}^{k} i_l = n$. In addition, consider independent random variables $U(l)$, $V(i,l)$ which are independent of I and such that $U(l)$ has m.g.f. $\mu \cdot \{M(z)\}^l \cdot \mathbf{1}$ and $V(i,l)$ has m.g.f. $\mathbf{e}(z) \cdot \{M(z)\}^{l-1} \cdot \mathbf{1}$. If one defines $W_I := U(I_0) + \sum_{l=1}^{K} V(l, I_l)$ then*

$$\|T_n - W_I\| \leq (1 - \ell_n). \tag{13}$$

As a numerical example, we select a stationary and homogenous Markov chain from [Erh99] with state space $S = \{1, \dots, 8\}$ and probability transition matrix

$$p(i,j) = \begin{cases} \frac{1-\beta q(i,T)}{1-q(i,T)} q(i,j) & , \quad i \in T^c, j \in T^c; \\ \beta q(i,j) & , \quad i \in T^c, j \in T; \\ q(i,j) & , \quad i \in T, j \in S; \end{cases} \tag{14}$$

where

$$q = \begin{pmatrix} 0.334 & 0.215 & 0.173 & 0.119 & 0.065 & 0.086 & 0.003 & 0.005 \\ 0.289 & 0.133 & 0.211 & 0.133 & 0.067 & 0.156 & 0.007 & 0.004 \\ 0.356 & 0.184 & 0.075 & 0.043 & 0.151 & 0.183 & 0.002 & 0.006 \\ 0.41 & 0.162 & 0.108 & 0.075 & 0.14 & 0.097 & 0.005 & 0.003 \\ 0.316 & 0.239 & 0.044 & 0.218 & 0.076 & 0.098 & 0.004 & 0.005 \\ 0.44 & 0.176 & 0.044 & 0.242 & 0.088 & 0 & 0.005 & 0.005 \\ 0.18 & 0.06 & 0.19 & 0.09 & 0.13 & 0.1 & 0.13 & 0.12 \\ 0.2 & 0.16 & 0.07 & 0.1 & 0.14 & 0.1 & 0.09 & 0.14 \end{pmatrix}.$$

The goal is to approximate the occupancy distribution of the set $T = \{8\}$ for various values of n and β. The parameter β controls transitions to T, which become rare for β small. Table 2 gives exact total variation distance errors for Normal [FS09], Poisson [BHJ92], and compound Poisson approximations [Erh99] as well as our approximation in (13). The compound Poisson approximation is a Pólya-Aeppli distribution.

The Poisson approximations are most effective when β is small. However, because $p(8,8) = 0.14$, regardless of β, visits to T may occur in clumps. In particular, the Pólya-Aeppli distribution is a more natural approximation to the occupancy distribution of T.

As shown in the table, approximation (13) gives one order of magnitude or more improvement over both Poisson approximations. Furthermore, it is clear that $n = 1000$ may be not large enough for an accurate Normal approximation to the occupancy distribution of T.

n	β	Normal	Poisson	Compound Poisson	Approximation in (13)
10	1	1.7E-2	1.4E-2	3.2E-3	3.8E-4
10	0.5	1.7E-2	7.0E-3	1.2E-3	1.5E-4
10	0.25	1.3E-2	3.6E-3	4.9E-4	6.9E-5
10	0.1	5.3E-3	1.4E-3	1.7E-4	2.7E-5
10	0.01	5.3E-4	1.4E-4	1.5E-5	2.6E-6
10	0.001	5.3E-5	1.4E-5	1.5E-6	2.6E-7
100	1	0.23	6.9E-2	9.7E-3	2.3E-4
100	0.5	0.22	5.2E-2	3.5E-3	1.6E-4
100	0.25	0.14	3.2E-2	1.3E-3	7.5E-5
100	0.1	2.0E-2	1.5E-2	3.1E-4	3.1E-5
100	0.01	5.2E-3	1.6E-3	1.6E-5	3.3E-6
100	0.001	5.3E-4	1.6E-4	1.5E-6	3.3E-7
1000	1	6.9E-2	7.0E-2	9.4E-3	2.1E-5
1000	0.5	9.0E-2	7.3E-2	4.9E-3	1.4E-5
1000	0.25	0.14	7.8E-2	2.7E-3	8.2E-6
1000	0.1	0.23	6.8E-2	9.6E-4	1.1E-5
1000	0.01	2.0E-2	1.5E-2	2.7E-5	1.8E-6
1000	0.001	5.2E-3	1.5E-3	1.7E-6	2.0E-7

Tab. 2: Total variation distance for approximations to the occupancy distribution of the set $T = \{8\}$ for the stationary chain described by (14). The compound Poisson approximation, given by [Erh99], is a Pólya-Aeppli distribution.

Acknowledgements

The authors are indebted to Emeritus Professor Eugene Seneta for providing a hardcopy of Doeblin's 1937 report, and for his helpful comments on an earlier draft of this manuscript.

References

[AGG90] R. Arratia, L. Goldstein, and L. Gordon. Poisson approximation and the Chen-Stein method. *Stat. Sci.*, 5(4):403–424, 1990.

[BC87] J. D. Biggins and C. Cannings. Markov renewal processes, counters and repeated sequences in Markov chains. *Adv. Appl. Prob.*, 19:521–545, 1987.

[BHJ92] A. D. Barbour, L. Holst, and S. Janson. *Poisson Approximation*. Oxford University Press, first edition, 1992.

[BK93] E. A. Bender and F. Kochman. The distribution of subword counts is usually normal. *Eur. J. Comb.*, 14(4):265–275, 1993.

[Che10] S. Chestnut. Approximating Markov chain occupancy distributions. Master's thesis, University of Colorado, The United States, April 2010.

[CT01] J. N. Corcoran and R. L. Tweedie. Perfect sampling of ergodic Harris chains. *Ann. Appl. Probab.*, 11(2):438–451, 2001.

[Dob56a] R. L. Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory Probab. Appl.*, 1(1):65–79, 1956.

[Dob56b] R. L. Dobrushin. Central limit theorem for nonstationary Markov chains. II. *Theory Probab. Appl.*, 1(4):329–383, 1956.

[Doe37] W. Doeblin. Le cas discontinu des probabilités en chaîne. *Publ. Fac. Sci. Univ. Masaryk (Brno)*, 236:1–13, 1937.

[Doe38] W. Doeblin. Exposé de la theorie des chaînes simple constantes de Markov à un nombre fini d'états. *Rev. Math. Union Interbalkan*, 2:77–105, 1938.

[Dur99] R. Durrett. *Essentials of stochastic processes*. Springer, first edition, 1999.

[Erh99] T. Erhardsson. Compound Poisson approximation for Markov chains using Stein's method. *Ann. Prob.*, 27:565–596, 1999.

[Fel68] W. Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons, third edition, 1968.

[Fro12] G. Frobenius. Über Matrizen aus nicht negativen Elementen. *Sitz.-Ber. Akad. Wiss., Phys-Math Klasse, Berlin*, pages 456–477, 1912.

[FS09] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, first edition, 2009.

[GL81] H. U. Gerber and S.-Y. R. Li. The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain. *Stoch. Proc. Appl.*, 11(1):101–108, 1981.

[Gri75] D. Griffeath. Uniform coupling of non-homogeneous Markov chains. *J. Appl. Probab.*, 12(4):753–762, 1975.

[Haj58] J. Hajnal. Weak ergodicity in nonhomogeneous Markov chains. *Proc. Cambridge. Philos. Soc.*, 54:233–246, 1958.

[Ios72] M. Iosifescu. On two recent papers on ergodicity in nonhomogeneous Markov chains. *Ann. Math. Statist.*, 43(5):1732–1736, 1972.

[KLW$^+$10] R. Kennedy, M. E. Lladser, Z. Wu, C. Zhang, M. Yarus, H. De Sterck, and R. Knight. Natural and artificial RNAs occupy the same restricted region of sequence space. *RNA*, 16(2):280–289, 2010.

[KLYK08] R. Kenney, M. E. Lladser, M. Yarus, and R. Knight. Information, probability, and the abundance of the simplest RNA active sites. *Front. Biosci.*, 13:6060–71, 2008.

[Lin71] M. Lin. Mixing for Markov operators. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 19:231–242, 1971.

[Lin77] T. Lindvall. A probabilistic proof of Blackwell's renewal theorem. *Ann. Probab.*, 5(3):482–485, 1977.

[Lla07] M. E. Lladser. Minimal Markov chain embeddings of pattern problems. In *Proceedings of the 2007 Information Theory and Applications Workshop*, University of California, San Diego, 2007.

[Lla08] M. E. Lladser. Markovian embeddings of general random strings. In *2008 Proceedings of the Fifth Workshop on Analytic Algorithmics and Combinatorics*, pages 183–190, San Francisco, California, 2008. SIAM.

[Mar06] A. A. Markov. Extension of the law of large numbers to dependent quantities (in Russian). *Izvestiia Fiz.-Matem. Obsch. Kazan Univ.*, 15:135–156, 1906. (2nd Ser.).

[MG98] D. J. Murdoch and P. J. Green. Exact sampling from a continuous state space. *Scand. J. Stat.*, 25(3):483–502, 1998.

[Møl99] J. Møller. Perfect simulation of conditionally specified models. *J. R. Statist. Soc. B*, 61(1):251–264, 1999.

[Nic03] P. Nicodème. Regexpcount, a symbolic package for counting problems on regular expressions and words. *Fund. Inform.*, 56(1-2):71–88, 2003.

[NSF02] P. Nicodème, B. Salvy, and P. Flajolet. Motif statistics. *Theoret. Comput. Sci.*, 287(2):593–617, 2002.

[Paz70] A. Paz. Ergodic theorems for infinite probabilistic tables. *Ann. Math. Statist.*, 41(2):539–550, 1970.

[Per07] O. Perron. Über Matrizen. *Math. Ann.*, pages 248–263, 1907.

[Pit74] J. W. Pitman. Uniform rates of convergence for Markov chain transition probabilities. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 29:193–227, 1974.

[RS07] E. Roquain and S. Schbath. Improved compound Poisson approximation for the number of occurrences of any rare word family in a stationary Markov chain. *Adv. in Appl. Probab.*, 39(1):128–140, 2007.

[Sen73a] E. Seneta. *Non-negative matrices*. John Wiley & Sons, first edition, 1973.

[Sen73b] E. Seneta. On the historical development of the theory of finite inhomogeneous Markov chains. *Proc. Cambridge Philos. Soc.*, 74:507–513, 1973.

[Sen93] E. Seneta. Applications of ergodicity coefficients to homogeneous Markov chains. In *Doeblin and modern probability*, volume 149 of *Contemp. Math*, pages 189–199. Amer. Math. Soc., Providence, RI, 1993.

[Tho90] H. Thorisson. The classical coupling, a refinement. *Teor. Veroyatnost. i Primenen.*, 35(4):809–817, 1990.