Doeblin's ergodicity coefficient: lower-complexity approximation of occupancy distributions

M. E. Lladser, joint work with S. Chestnut
Department of Applied Mathematics
University of Colorado - Boulder

Notation.

S is a **finite** set of states

$X = (X_t)_{t \geq 0}$ is a first-order homogeneous Markov chain with:

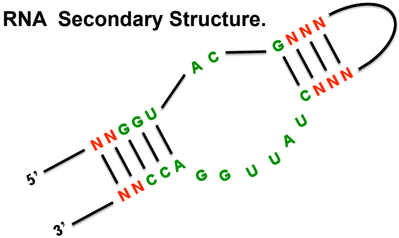
- **initial distribution** μ
- **probability transition matrix** $p = (p_{i,j})_{i,j \in S}$
- **stationary distribution** π when irreducible

Object of interest.

The **(n-th step) occupancy distribution** of a set $T \subset S$:

$$T_n \stackrel{def}{=} \#\,(\text{visits to } T \text{ in first } n\text{-transitions})$$
$$= \sum_{t=1}^{n} [\![X_t \in T]\!], \text{ where } [\![A]\!] \text{ is the indicator of } A$$

Applications of occupancy distributions.



Pattern. 12GGUACG345★5'4'3'CUAUUGGACC2'1'

Figure. (i) What's the probability the pattern occurs somewhere in a random RNA of length-100? (ii) Given that the pattern does not occur, what's the probability that *GGUACG* occurs 7-times?

Embedding Technique.
[GERBER-LI'81, BIGGINS-CANNINGS'87, BENDER-KOCHMAN'93]

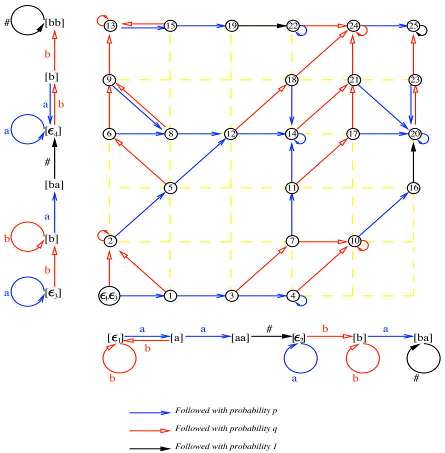


Figure. Markov chain that keeps tracks of the joint presence/absence of the pattern $1a\#b1$ in an i.i.d. $\{a,b\}$-text [LLADSER-BETTERTON-KNIGHT'08]

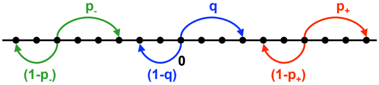
Embedding of non-Markovian sequences. [LLADSER'08]

Consider the $\{0,1\}$-valued stochastic sequence

$$X_{n+1} \overset{d}{=} \begin{cases} \text{Ber}(p_+) & \text{, if } \frac{1}{n}\sum_{i=1}^{n} X_i > g; \\ \text{Ber}(q) & \text{, if } \frac{1}{n}\sum_{i=1}^{n} X_i = g; \\ \text{Ber}(p_-) & \text{, if } \frac{1}{n}\sum_{i=1}^{n} X_i < g; \end{cases} \quad \text{with } g = \frac{1}{1+r/l} \text{ and } \gcd(r,l)=1$$

Theorem.

If $\mathcal{L}$ is a **regular pattern** and $\mathcal{Q}$ the set of states of any **deterministic finite automaton** that recognizes $\mathcal{L}$ then, there is homogeneous Markov chain with state space $\mathbb{Z} \times \mathcal{Q}$ which keeps tracks of all the prefixes of the infinite sequence X that belong to $\mathcal{L}$. Furthermore, the projection of the chain into $\mathbb{Z}$ is also a Markov chain, with transition probabilities:



Pros & Cons in Literature.

$$T_n = \#\,(\text{visits to } T \text{ in first } n\text{-transitions}) = \sum_{t=1}^{n} [\![X_t \in T]\!]$$

Method	Formulation	Assumption	Weakness
Exact via recursions or operators [DURRETT'99, FLAJOLET-SEDGEWICK'09]	$\mathbb{E}(x^{T_n}) = \mu \cdot p_x^n \cdot \mathbf{1}$	$\emptyset$	Complexity $O(n^2\|S\|^2)$
Normal Approx. [BENDER-KOCHMAN'93, RÉGNIER-SZPANKOWSKI'98, NICODÈME-SALVY-FLAJOLET'02]	$\sum_{n=0}^{\infty} y^n \cdot \mathbb{E}(x^{T_n}) = \mu \cdot (\mathbb{I} - y \cdot p_x)^{-1} \cdot \mathbf{1}$	Irreducibility, aperiodicity	$O\left(\frac{1}{\sqrt{n}}\right)$-rate of convergence
Poisson Approx. [ALDOUS'88, BARBOUR-HOLST-JANSON'92]	$T_n \overset{d}{\approx} Poisson(n \cdot \pi(T))$	Stationarity	Ignores clumps of visits to T
Compound Poisson Approx. [ERHARDSSON'99, ROQUAIN-SCHBATH'07]	$T_n \overset{d}{\approx} CPoisson(n\lambda_1, n\lambda_2, \ldots)$; $\lambda_i = [x^k]\,\nu \cdot (\mathbb{I} - q_x)^{-1} \cdot r$	Stationarity	Needs atom s s.t. $\mathbb{P}_\pi(\tau_T < \tau_s) \ll 1$

Challenge.

To approximate the distribution of T_n when n is perhaps **too large for exact calculations and too small to rely on the Normal approximation**, ...

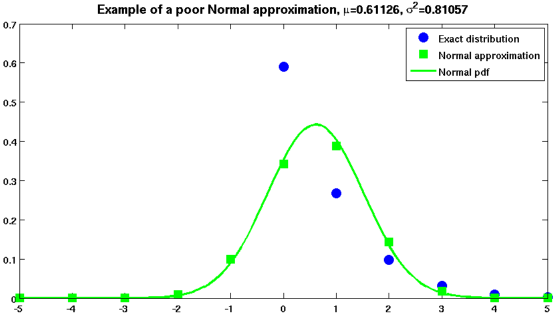


Figure. Normal approximation for a stationary chain considered in [ERHARDSSON'99] with $S = \{1, \ldots, 8\}$, $T = \{8\}$ and $n = 1000$

Motivations.

Challenge.

To approximate the distribution of T_n when n is perhaps too large for exact calculations and too small to rely on the Normal approximation, **and without assuming that X is stationary**

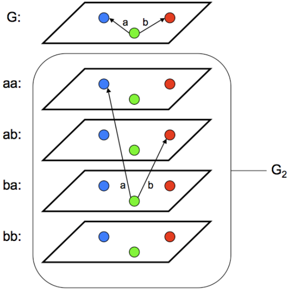


Figure. Second-order automaton associated with automaton G on top [NICODÈME-SALVY-FLAJOLET'02, LLADSER'07]

Addressing the challenge.

All the complexity associated with approximating the distribution of

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

is due to the dependence between X_t and X_{t-1}, for $1 \leq t \leq n$. Overlooking this dependence is naive, however, the extent of dependence could be reduced if one could **guess at random times where the chain is located**. To achieve this, we assume the following

Standing hypothesis.

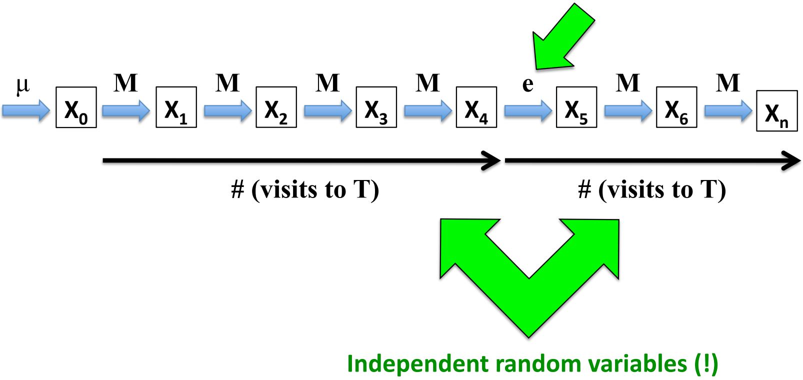
There is $\lambda > 0$ and stochastic matrices E and M s.t. $p = \lambda \cdot E + (1 - \lambda) \cdot M$, where all rows of E are identical to certain probability vector $\mathbf{e}$

- p satisfies **Doeblin's condition** [Doeblin'40]: $p^m(i,j) \geq \lambda \cdot \mathbf{e}(j)$, with $m = 1$
- One can simulate from π exactly without computing it beforehand, using the **multi-gamma coupling** [Murdoch-Green'98, Møller'99, Corcoran-Tweedie'01]

Approximating the distribution of T_n, with $n = 7$.

Standing hypothesis.

$(\exists\lambda > 0): p = \lambda \cdot E + (1 - \lambda) \cdot M$, where all rows of E are identical to $\mathbf{e}$

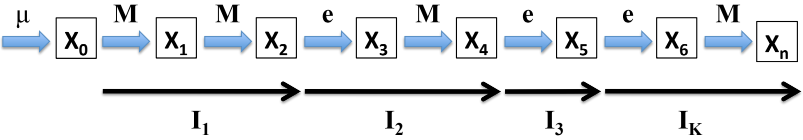


Approximating the distribution of T_n, with $n = 7$.

Standing hypothesis.

$(\exists\lambda > 0): p = \lambda \cdot E + (1-\lambda) \cdot M$, where all rows of E are identical to **e**

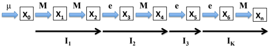
A perhaps more likely scenario (!)



In which case the largest transfer matrix exponent to consider is 2 rather than n

Standing hypothesis.

$(\exists \lambda > 0): p = \lambda \cdot E + (1 - \lambda) \cdot M$, where all rows of E are identical to **e**



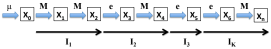
Heuristic.

The **largest transfer-matrix power** to consider is

$$L_n = \max_{i=1,\ldots,K} I_i,$$

which **concentrates** around $\frac{-\ln(\lambda n)}{\ln(1-\lambda)}$ [Feller'68, Arratia-Goldstein-Gordon'90, Flajolet-Sedgewick'09]. **Accurate approximations** to the distribution of T_n should **follow by considering chains of duration** $m = \Theta\big(\ln(n)\big)$ **instead of** n

Approximating the distribution of T_n, with $n = 7$.



Theorem [Chestnut-Lladser'10].

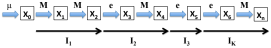
If $W_{n,m}$ is the random number of visits to T when $L_n \leq m$ then

$$\|T_n - W_{n,m}\| \quad \leq \quad \mathbb{P}[L_n > m] \quad \sim O(n^{1-c}), \text{ when } m = \frac{c \cdot \ln(\lambda n)}{\ln(1/(1-\lambda))}$$

Rate of convergence.

$c = 3/2$ matches the rate of convergence of the Normal approximation (!)

Approximating the distribution of T_n, with $n = 7$.



Theorem [Chestnut-Lladser'10].

If $W_{n,m}$ is the random number of visits to T when $L_n \leq m$ then

$$\|T_n - W_{n,m}\| \quad \leq \quad \mathbb{P}[L_n > m] \quad \sim O(n^{1-c}), \text{ when } m = \frac{c \cdot \ln(\lambda n)}{\ln(1/(1-\lambda))}$$

Numerical Implementation.

The combinatorial class of coin flips with M-runs of length $\leq m$ is described by the **regular expression**:

$$(\epsilon + \mu\{M, \ldots, M^m\}) \times (\mathbf{e}\{M, \ldots, M^m\})^*,$$

implying that $\sum_{k \geq 0} \mathbb{E}(x^{W_{k,m}}) y^k$ is **rational** and **computable from** $(\mu \cdot M_x^l \cdot \mathbf{1}) y^l$ and $(\mathbf{e}_x \cdot M_x^l \cdot \mathbf{1}) y^{l+1}$, with $l = 0, \ldots, m$

Small numerical example.

n	δ	Normal approximation	Poisson approximation	Compound Poisson approximation	Our approximation
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

Table. Errors in total variation distance for stationary chains considered in [ERHARDSSON'99], where $S = \{1, \ldots, 8\}$ and $T = \{8\}$. The parameter δ controls transitions into T, which are rare for δ small

Looking back ...

Standing hypothesis.

$(\exists\lambda > 0)$: $p = \lambda \cdot E + (1-\lambda)\cdot M$, where all rows of E are identical to $\mathbf{e}$

To aim at the best approximation, choose:

$$\max\left\{\lambda :\ \exists E\,\exists M :\ p = \lambda \cdot E + (1-\lambda)\cdot M\right\} = \sum_j \min_i p(i,j).$$

i.e. the optimal λ is **Doeblin's ergodicity coefficient** [DOEBLIN'37] associated with p:

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

Several other ergodicity coefficients have been introduced in the literature [MARKOV'906, DOBRUSHIN'56, HAJNAL'58, SENETA'73+'93] e.g. the **Markov-Dobrushin ergodicity coefficient** is:

$$\beta(p) \stackrel{def}{=} 1 - \max_{i,j} \|p(i,\cdot) - p(j,\cdot)\|_{tvd} \qquad (\geq \alpha(p))$$

Ok! ... what if $\alpha(p) = 0$?

In the aperiodic and irreducible setting:

$$\lim_{k\to\infty} p^k = \Pi \quad \Longrightarrow \quad \lim_{k\to\infty} \alpha(p^k) = 1$$

It is well-known that the **Markov-Dobrushing coefficient is sub-multiplicative** [Dobrushin'56, Paz'70, Iosifescu'72, Griffeath'75]:

$$(\forall p, q \in \mathcal{P}): \quad \big(1 - \beta(pq)\big) \leq \big(1 - \beta(p)\big) \cdot \big(1 - \beta(q)\big)$$

Exploiting that

$$\begin{aligned} p &= \alpha(p) \cdot E_1 + \big(1 - \alpha(p)\big) \cdot M_1 \\ q &= \alpha(q) \cdot E_2 + \big(1 - \alpha(q)\big) \cdot M_2 \end{aligned}$$

we obtain:

Theorem [Chestnut-Lladser'10?].

$(\forall p, q \in \mathcal{P}): \quad \big(1 - \alpha(pq)\big) \leq \big(1 - \alpha(p)\big) \cdot \big(1 - \alpha(q)\big)$

An unexpected consequence for non-homogeneous chains.

Doeblin's characterization of weak-ergodicity (1937).

For a sequence of stochastic matrices $(p_k)_{k\geq 0}$ the following are equivalent:

- $(\forall m \geq 0)(\forall i,j,s \in S) : \lim_{n\to\infty} \left| \left(\prod_{k=m}^{n} p_k\right)(i,s) - \left(\prod_{k=m}^{n} p_k\right)(j,s) \right| = 0$
- there exists a strictly increasing sequence of positive integers $(n_k)_{k\geq 0}$ such that: $\sum_{k=0}^{\infty} \alpha\left(\prod_{i=n_k}^{n_{k+1}-1} p_i\right) = +\infty$

Similar characterizations but based on the β-coefficient were provided by Hajnal (1958), Paz (1970), and Iosifescu (1972), with increasing level of generality. Seneta (1973) proved Doeblin's characterization using various relationships between $\alpha(p)$, $\beta(p)$, and:

$$\gamma_1(p) \overset{def}{=} \max_j \min_i p(i,j), \text{ and } \gamma_2(p) \overset{def}{=} 1 - \max_s \max_{i,j} |p(i,s) - p(j,s)|$$

Using the sub-multiplicative inequality, we can now prove Doeblin's characterization in an elementary and self-contained way!

Main Reference.

[$\star$] OCCUPANCY DISTRIBUTIONS IN MARKOV CHAINS VIA DOEBLIN'S ERGODICITY COEFFICIENT. S. Chesnut, M. E. Lladser. *Discrete Mathematics and Theoretical Computer Science Proceedings.* AM, 79-92 (2010).

... Thank you!

A first-principles proof.

Doeblin's characterization of weak-ergodicity.

$(p_k)_{k\geq 0}$ is weakly-ergodic iff there exists a strictly increasing sequence of positive integers $(n_k)_{k\geq 0}$ such that: $\sum_{k=0}^{\infty} \alpha\Big(\prod_{i=n_k}^{n_{k+1}-1} p_i\Big) = +\infty$

Fix $m \geq 0$ and let $\alpha_n = \alpha\left(\prod_{k=m}^{n} p_k\right)$. Notice:

$$\Big(\prod_{k=m}^{n} p_k\Big)(i,s) - \Big(\prod_{k=m}^{n} p_k\Big)(j,s) = (1-\alpha_n)\cdot\big(M_n(i,s) - M_n(j,s)\big), \text{ with } \alpha(M_n) = 0$$

Proof of sufficiency [Chestnut-Lladser'10].

Using the sub-multiplicative property:

$$(1-\alpha_n) \leq \prod_{k\in K_n}\left\{1 - \alpha\Big(\prod_{i=n_k}^{n_{k+1}-1} p_i\Big)\right\} \leq \exp\left\{-\sum_{k\in J_n}\alpha\Big(\prod_{i=n_k}^{n_{k+1}-1} p_i\Big)\right\}$$

A first-principles proof.

Doeblin's characterization of weak-ergodicity.

$(p_k)_{k\geq 0}$ is weakly-ergodic iff there exists a strictly increasing sequence of positive integers $(n_k)_{k\geq 0}$ such that: $\sum_{k=0}^{\infty} \alpha\Big(\prod_{i=n_k}^{n_{k+1}-1} p_i\Big) = +\infty$

Fix $m \geq 0$ and let $\alpha_n = \alpha\left(\prod_{k=m}^{n} p_k\right)$. Notice:

$$\Big(\prod_{k=m}^{n} p_k\Big)(i,s) - \Big(\prod_{k=m}^{n} p_k\Big)(j,s) = (1-\alpha_n)\cdot\big(M_n(i,s) - M_n(j,s)\big), \text{ with } \alpha(M_n) = 0$$

Proof of necessity [Chestnut-Lladser'10].

It suffices to prove that $(\alpha_n)_{n\geq 0}$ has a subsequence that converges to 1. By contradiction, if one assumes otherwise then

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

However, this is not possible because M_n has a zero in each column and M_n is a stochastic matrix