On Martingales, Markov Chains and Concentration

David Aldous

28 June 2018

When I was a Ph.D. student (1974-77) Markov Chains were a long-established and still active topic in Applied Probability; and Martingales were a long-established and still active topic in Theoretical Probability. But (according to memory) there wasn't much connection between those topics. Maybe martingales were a potentially useful tool for studying Markov Chains, but were they actually being used?

Here are the results of a MathSciNet search on "year = 1977" and "anywhere = martingale and Markov chain".

AMERICAN MATHEMATICAL SOCIETY
MathSciNet
Mathematical Reviews
ISSN 2167-5163

Publications results for "(Anywhere=(Markov chain) AND Anywhere=(martingale) AND Publication Type=(Journals)) AND pubyear=1977 "

MR0482943 (58 #2978) Reviewed
Bolthausen, E.
On rates of convergence in a random central limit theorem and in the central limit theorem for Markov chains.
Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **38** (1977), no. 4, 279–286.
60F05 (60J10)
UC-eLinks

Let $\{X_n, n = 0, 1, \cdots\}$ be a positive recurrent irreducible Markov chain with countable state space I, and assume that a map $\varphi: I \to \mathbf{R}$ (reals) is given. Fix a state $a \in I$ and define $\tau := \inf\{n \geq 1: X_n = a\}$ and $\varphi_\tau := \sum_{n=1}^{\tau} \varphi(X_n)$. Using renewal theory, optional stopping results for martingales and some well-known Berry-Esseen type results for i.i.d. summands, it is proved, given specified constants μ and σ, $F_n(t) := P_a(\sum_1^n \varphi(X_k) - n\mu \leq n^{1/2}\sigma t)$ and Φ the standard normal distribution, that $E_a|\tau|^{4\vee c} < \infty$ and $E_a|\phi_\tau|^c < \infty$ for some $c \geq 3$ imply

$$\|F_n - \Phi\| = o(n^{-1/3+\delta}) \quad \text{for all} \delta > 1/(6(c+1)).$$

Reviewed by Roy V. Erickson

Fast forward to 1987

- ☐ Stochastic Process. Appl. (1)

Year

- ☐ 1987 (7)

MR0943579 Reviewed Aalen, Odd O. Dynamic modelling and causality. *Scand. Actuar. J.* 1987, no. 3-4, 177–190. 62P10 (60G44 62M10 62P05)
UC-eLinks

MR0926552 Reviewed Moore, Marc Inférence statistique dans les processus stochastiques: aperçu historique. (French) [Statistical inference in stochastic processes: a historical overview] *Canad. J. Statist.* 15 (1987), no. 3, 185–207. (Reviewer: W. A. O'N. Waugh) 62-02 (01A60 60-02 62M99)
UC-eLinks

MR0925936 Reviewed Aldous, David; Shepp, Larry The least variable phase type distribution is Erlang. *Comm. Statist. Stochastic Models* 3 (1987), no. 3, 467–473. 60J27 (90B22)
UC-eLinks

MR0917251 Reviewed Bierth, K.-J. An expected average reward criterion. *Stochastic Process. Appl.* 26 (1987), no. 1, 123–140. (Reviewer: Uriel G. Rothblum) 90C39 (90C40)
UC-eLinks

MR0900114 Reviewed Sonin, I. M. A theorem on separation of jets and some properties of random sequences. *Stochastics* 21 (1987), no. 3, 231–249. (Reviewer: Robert Kertz) 60J10 (60F99 60G44)
UC-eLinks

MR0889799 Reviewed Fayolle, Guy; Iasnogorodski, Rudolph Criteria for the nonergodicity of stochastic processes: application to the exponential back-off protocol. *J. Appl. Probab.* 24 (1987), no. 2, 347–354. (Reviewer: Linn I. Sennott) 60G40 (60K99 68M10 94A05)
UC-eLinks

MR0877725 Reviewed Rosenkrantz, Walter A. Approximate counting: a martingale approach. *Stochastics* 20 (1987), no. 2, 111–120. (Reviewer: S. Csibi) 68R99 (60G42 68Q25)
UC-eLinks

I will first talk about Aldous-Shepp (1987) and then about some recent work.

Consider the rather trivial $n+1$-state continuous-time chain (X_t) with states $0, 1, \ldots, n$ and transition rates

$$q_{i,i-1} = 1, \ 1 \leq i \leq n.$$

The hitting time $T = T_{n,0}$ from n to 0 has relative variability

$$\mathrm{var}(T)/(\mathbb{E}T)^2 = 1/n.$$

It is natural to conjecture that this is the smallest possible value for any hitting time of any $n+1$-state continuous-time chain, and Aldous-Shepp (1987) gave a short proof, based on the following general identity.

Lemma

Let T be a first hitting time and write $h(i) := \mathbb{E}_i T$.
Then $\mathrm{var}_i \ T = \mathbb{E}_i \sum_{s \le T} (h(X_s) - h(X_{s-}))^2 = \mathbb{E}_i \int_0^T a(X_t) \ dt$
where $a(i) := \sum_j q_{ij} \ (h(i) - h(j))^2$.

Now w.l.o.g. label states as $0, 1, \ldots, n$, consider the hitting time on state 0, and re-order so that $0 = h(0) < h(1) \le h(2) \le \ldots \le h(n)$. Easy to see that for any path $i_0, i_1, \ldots$ from m to 0 we have

$$\sum_u (h(i_u) - h(i_{u-1}))^2 \ge \sum_{i=1}^m (h(i) - h(i-1))^2.$$

So

$$\begin{aligned}
\mathrm{var}_m T_0 &\ge \sum_{i=1}^m (h(i) - h(i-1))^2 && \text{(first equality in Lemma)} \\
&\ge m^{-1} \left(\sum_{i=1}^m (h(i) - h(i-1)) \right)^2 && \text{(Cauchy-Schwarz)} \\
&\ge m^{-1} h^2(m) \ge n^{-1} (\mathbb{E}_m T_0)^2.
\end{aligned}$$

This result is mentioned in treatments of "phase-type" distributions but neither Larry nor I pursued the topic further.

The "theory" way to understand the lemma is that $t \to \mathbb{E}(T|X_s, s \le t)$ must be a martingale, and in fact it is

$$M_t := \mathbb{E}(T|X_s, s \le t) = h(X_{t \wedge T}) + t \wedge T \qquad (1)$$

for $h(i) := \mathbb{E}_i T$. Next, M_t^2 has a Doob-Meyer decomposition into a martingale Q_t and a predictable process, and the decomposition is

$$M_t^2 - M_0^2 = Q_t + \int_0^t a(X_s) ds, \ \ t \le T$$

for

$$a(i) := \sum_j q_{ij} \ (h(i) - h(j))^2.$$

From these ingredients and optional sampling we get a pair of general identities.

Lemma

Let T be a first hitting time within a continuous-time chain, and

$$h(i) := \mathbb{E}_i T.$$

Then

$$\mathbb{E}_i T = \mathbb{E}_i \int_0^T b(X_t)\ dt, \quad \mathrm{var}_i\ T = \mathbb{E}_i \int_0^T a(X_t)\ dt$$

where

$$a(i) := \sum_j q_{ij}\ (h(i) - h(j))^2$$

$$b(i) := \sum_j q_{ij}(h(i) - h(j)).$$

These are curiously hard to find in Applied Probability textbooks, and indeed are not helpful for explicit calculations in a specific model.

As background to my second topic, recall the **method of bounded differences;** for a RV Z of the form

$$Z = f(\xi_1, \dots \xi_n); \quad \text{for independent } (\xi_i)$$

where f has the property

$$|f(\mathbf{x}) - f(\mathbf{x}')| \leq 1 \text{ whenever } \mathbf{x},\ \mathbf{x}' \text{ differ in only one coordinate}$$

we have

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq \lambda n^{1/2}) \leq 2 \exp(-\lambda^2/2).$$

Basic example of a general **concentration inequality** – a key point is that one can bound a difference $|Z - \mathbb{E}Z|$ even when you don't know $\mathbb{E}Z$.

I will describe a rather different CI, which holds for Markov chains with a special property.

Consider a finite-state continuous-time Markov chain and a hitting time $T = T_A$ on some subset A of states, and suppose

$$h(i) := \mathbb{E}_i T < \infty \ \forall i.$$

Neither "finite" nor "continuous" is actually important here. What is important is the next assumption:

$$h(j) \le h(i) \text{ for each possible transition } i \to j. \tag{2}$$

This is a very restrictive assumption – not obvious that **any** "interesting" chain satisfies this assumption.

Proposition

Under condition (2), for any initial state,

$$\frac{\mathrm{var}\ T}{\mathbb{E} T} \le \max\{h(i) - h(j) : \ i \to j \text{ a possible transition}\}.$$

We can rewrite this as

$$\kappa := \max\{h(i) - h(j) : \ i \to j \text{ a possible transition}\}$$

$$\frac{\mathrm{s.d.}(T)}{\mathbb{E}T} \leq \sqrt{\frac{\kappa}{\mathbb{E}T}}$$

where the right side is always ≤ 1. So we get a weak concentration inequality if $\kappa/\mathbb{E}T$ is small.

The Proposition is an immediate corollary of the previous lemma [show]

Are there any interesting chains where our "strong monotonicity" condition holds?

Our applications are all in the context of chains (Z_t) whose states are subsets S of a given discrete space and whose transitions are of the form $S \to S \cup \{v\}$. In words "increasing set-valued processes".

And our applications use hitting times T of the form

$$T := \inf\{t : Z_t \supseteq B \text{ for some } B \in \mathcal{B}\}$$

for a specified collection $\mathcal{B}$ of subsets B.

In applications we bound
$\kappa := \max\{h(i) - h(j) : \ i \to j \text{ a possible transition}\}$ by using some natural coupling of the processes started from i and from j.

The rest of the talk is 3 examples which fit this context. The first two are straightforward to implement.

Example 1: a general growth process (Z_t) on the lattice $\mathbb{Z}^2$.
The states are finite vertex-sets S, the possible transitions are $S \to S \cup \{v\}$ where v is a vertex adjacent to S. For each such transition, we assume the transition rates are bounded above and below:

$$0 < c_* \leq q(S, S \cup \{v\}) \leq c^* < \infty. \tag{3}$$

Initially $Z_0 = \{\mathbf{0}\}$, where $\mathbf{0}$ denotes the origin. The "monotonicity" condition we impose is that these rates are increasing in S:

$$\text{if } v, v' \text{ are adjacent to } S \text{ then } q(S, S \cup \{v\}) \leq q(S \cup \{v'\}, S \cup \{v, v'\}) \ . \tag{4}$$

Note that we do not assume any kind of spatial homogeneity.

Proposition

Let B be an arbitrary subset of vertices $\mathbb{Z}^2 \setminus \{\mathbf{0}\}$, and consider $T := \inf\{t : Z_t \cap B$ is non-empty.$\}$. Under assumptions (3, 4),

$$\mathrm{var}\ T \leq \mathbb{E}T/c_*.$$

Proof. Condition (4) allows us to couple versions (Z'_t, Z''_t) of the process starting from states $S' \subset S''$, such that in the coupled process we have $Z'_t \subseteq Z''_t$ for all $t \geq 0$. In particular, $h(S) := \mathbb{E}_S T$ satisfies the monotonicity condition (2). To deduce the result from Lemma 1 we need to show that, for any given possible transition $S_0 \to S_0 \cup \{v_0\}$, we have

$$h(S_0) \leq h(S_0 \cup \{v_0\}) + 1/c_*. \quad (5)$$

Now by running the process started at S_0 until the first time T^* this process contain v_0, and then coupling the future of that process to the process started at $S_0 \cup \{v_0\}$, we have $h(S_0) \leq \mathbb{E}_{S_0} T^* + h(S_0 \cup \{v_0\})$. And $\mathbb{E}_{S_0} T^* \leq 1/c_*$ by (3), establishing (5).

Two recent arXiv posts deal with specific models of such non-homogeneous growth processes.

Asymptotic behavior of the Eden model with positively homogeneous edge weights by Bubeck and Gwynne

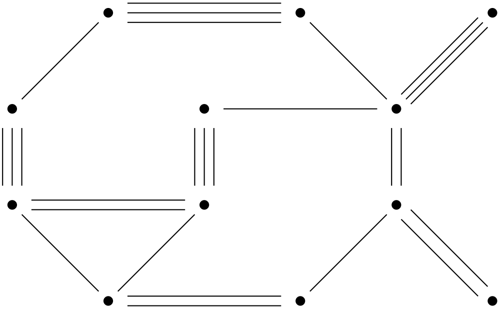
Nucleation and growth in two dimensions by Bollobas et al.

Example 2: A multigraph process
(This is one aspect of broad study in Aldous-Li (2018) *A Framework for Imperfectly Observed Networks*).

Take a **network** – a finite connected graph $(\mathbf{V}, \mathbf{E})$ with edge-weights $\mathbf{w} = (w_e)$, where $w_e > 0 \ \forall e \in \mathbf{E}$.

Define a multigraph-valued process as follows. Initially we have the vertex-set $\mathbf{V}$ and no edges. For each vertex-pair $e = (vy) \in \mathbf{E}$, edges vy appear at the times of a Poisson (rate w_e) process, independent over $e \in \mathbf{E}$.

So at time t the state of the process, Z_t say, is a multigraph with $N_e(t) \geq 0$ copies of edge e, where $(N_e(t), \ e \in \mathbf{E})$ are independent Poisson(tw_e) random variables.



Background motivation: model of imperfectly observed network. How well can we estimate some aspect of the true network from observed multigraph?

Leads to a huge range of questions; one basic issue is to say something about connectivity properties of true network. This must relate to connectivity properties of the observed multigraph.

We study how long until Z_t has various connectivity properties. Specifically, consider

- $T'_k = \inf\{t : Z_t \text{ is } k\text{-edge-connected}\}$
- $T_k = \inf\{t : Z_t \text{ contains } k \text{ edge-disjoint spanning trees.}\}$

Here we regard the $N_e(t)$ copies of e as disjoint edges. Remarkably, Lemma 1 enables us to give a simple proof of a "weak concentration" bound which does not depend on the underlying weighted graph.

Proposition

$$\frac{\mathrm{s.d.}(T_k)}{\mathbb{E}\, T_k} \leq \frac{1}{\sqrt{k}}, \quad k \geq 1.$$

Via a continuization device, the same bound holds in the discrete-time model where edges e arrive IID with probabilities proportional to w_e.

We conjecture that some similar result holds for T'_k. But proving this by our methods would require some structure theory (beyond Menger's theorem) for k-edge-connected graphs, and it is not clear whether relevant theory is known.

Proof. Here the states S are multigraphs over $\mathbf{V}$, and $h(S)$ is the expectation, starting at S, of the time until the process contains k edge-disjoint spanning trees. Monotonicity property is clear. What are the possible values of $h(S) - h(S \cup \{e\})$, where $S \cup \{e\}$ denotes the result of adding an extra copy of e to the multigraph S?

Consider the "min-cut" over proper subsets $S \subset \mathbf{V}$

$$\gamma := \min_S w(S, S^c)$$

where $w(S, S^c) = \sum_{v \in S, y \in S^c} w_{vy}$. Because a spanning tree must have at least one edge across the min-cut,

$$\mathbb{E} T_k \geq k/\gamma. \tag{6}$$

On the other hand we claim

$$h(S) - h(S \cup \{e\}) \leq 1/\gamma.$$

Given this, Lemma 1 establishes the proposition.

$$\text{Claim :} \qquad h(S) - h(S \cup \{e\}) \leq 1/\gamma.$$

To prove this, take the natural coupling (Z_t, Z_t^+) of the processes started from S and from $S \cup \{e\}$, and run the coupled process until Z_t^+ contains k edge-disjoint spanning trees. At this time, the process Z_t either contains k edge-disjoint spanning trees, or else contains $k-1$ spanning trees plus two trees (regard as edge-sets $\mathbf{t}_1$ and $\mathbf{t}_2$) such that $\mathbf{t}_1 \cup \mathbf{t}_2 \cup \{e\}$ is a spanning tree. So the extra time we need to run (Z_t) is at most the time until some arriving edge links $\mathbf{t}_1$ and $\mathbf{t}_2$, which has mean at most $1/\gamma$. This establishes the Claim.

Perhaps the most interesting example – from *The Incipient Giant Component in Bond Percolation ...* (2016).

Example 3: Bond percolation on a general network.

An edge e of weight w_e becomes **open** *at an Exponential(w_e) random time.*

In this process we can consider

$$C(t) = \text{max size (number of vertices) in a connected component of open edges at time } t.$$

And consider "**emergence of the giant component**". Studied extensively on many non-random and specific models of random networks. Can we say anything about (almost) arbitrary networks?

Traditional setting: number of vertices $n \to \infty$ asymptotics.

Suppose (after time-scaling) there exist constants $\delta > 0, K < \infty$ such that

$$\lim_n \mathbb{E} C_n(\delta)/n = 0; \quad \lim_n \mathbb{E} C_n(K)/n > 0. \tag{7}$$

In the language of random graphs, this condition says a *giant component* emerges (with non-vanishing probability) at some random time of order 1.

Proposition

Given a sequence of networks satisfying (1), there exist constants $\tau_n \in [\delta, K]$ such that, for every sequence $\varepsilon_n \downarrow 0$ sufficiently slowly, the random times

$$T_n := \inf\{t : C_n(t) \geq \varepsilon_n n\}$$

satisfy

$$T_n - \tau_n \to_p 0.$$

The Proposition asserts, informally, that the "incipient" time at which the giant component starts to emerge is deterministic to first order.

The proof of the Proposition is more complicated (we need to allow rare "bad" transitions) and less explicit (use a compactness argument) – but only 2-3 pages. See *The Incipient Giant Component in Bond Percolation . . .* (2016),

Why care about this kind of result?
Bond percolation can be re-interpreted as the SI epidemic. The Proposition can be re-interpreted as saying that, in the SI epidemic on an arbitrary network with an "infectiousness" parameter λ, there is always a critical value λ_{crit} such that, starting with $\Omega(1)$ but $o(n)$ infectives,
($\lambda < \lambda_{crit}$): w.h.p. $o(n)$ ever infected
($\lambda > \lambda_{crit}$): w.h.p. $\Omega(n)$ ever infected.

Would like to extend to more realistic SIR models, but needs different arguments.

Just for fun, a math example;
Take vertices as integers $1, 2, 3, \ldots, N$ and edge-weights

$$w_{ij} = \text{g.c.d.}(i, j)$$

with normalization $1/(N \log N)$. Here are 6 realizations of $C_N(\cdot)$ for $N = 72,000$.