Finite Markov Information-Exchange processes

David Aldous

February 2, 2011

Markov Chains

The next few lectures give a brisk discussion of

- Basics: discrete- and continuous-time.
- Hitting times and mixing times.
- Three standard examples.
- Other examples.

Only occasional math arguments here, **but** when we use some technique later for FMIEs that parallels a technique for MC, we'll recall the MC argument then.

A discrete-time MC $(Z(t), t = 0, 1, 2, \ldots)$ is specified by its transition matrix $\mathbf{P} = (p_{ij})$. The t-step transition probability are

$$\mathbb{P}_i(Z(t) = j) \text{ are entries of } \mathbf{P}^t.$$

A continuous-time MC $(Z(t),\ 0 \leq t < \infty)$ is specified by its transition rate matrix $\mathcal{N} = (\nu_{ij})$, where given the off-diagonal entries we set

$$\nu_i := \sum_{j \neq i} \nu_{ij}, \quad \nu_{ii} := -\nu_i.$$

The time-t transition probability are

$$\mathbb{P}_i(Z(t) = j) \text{ are entries of } \exp(\mathcal{N} t).$$

If **irreducible** then there is a unique **stationary distribution** $\pi = (\pi_i)$ and

$$\mathbb{P}_i(Z(t) = j) \to \pi_j \text{ as } t \to \infty$$

holds always in continuous-time. If we can find a distribution $\pi = (\pi_i)$ such that

$$\pi_i \nu_{ij} = \pi_j \nu_{ji} \ \forall j \neq i$$

then π is the stationary distribution and the chain is called **reversible**.

In our FMIE setting we use a symmetric matrix $\mathcal{N}$ and so we have an associated continuous-time MC which is reversible and has **uniform** stationary distribution. We sometimes impose assume **regularity**:

$$\sum_{j \neq i} \nu_{ij} = 1 \ \forall i. \tag{1}$$

This restriction is loosely analogous to discrete-time random walk on a graph being restricted to a regular graph.
Consider the (continuous-time) quantity

$$z_{ij} = \int_{t=0}^{\infty} (\mathbb{P}_i(Z(t) = j) - \pi_j) \ dt \tag{2}$$

analogous to the discrete-time quantity

$$z_{ij} = \sum_{t=0}^{\infty} \left(p_{ij}^{(t)} - \pi_j \right) \tag{3}$$

which can be viewed as a generalized inverse of the singular matrix $I - \mathbf{P}$. The matrix of mean hitting times $\mathbb{E}_i T_j^{\text{hit}}$ can be expressed in terms of the matrix $\mathbf{Z}$.

Theorem (Mean hitting time formula)

Without assuming reversibility, $\mathbb{E}_i T_j^{hit} = (z_{jj} - z_{ij})/\pi_j$.

See RWG 2.2 for proof and detailed discussion. Here, let me observe three consequences.
Noting $\sum_j z_{ij} = 0 \ \forall i$ we get $\sum_j (\mathbb{E}_i T_j^{\text{hit}})\pi_j = \sum_j z_{jj}$ and in particular

Corollary (Random Target Lemma)

$\tau_{hit} := \sum_j (\mathbb{E}_i T_j^{hit})\pi_j$ *does not depend on* i

and so this particular statistic τ_{hit} is the mathematically natural way to summarize the matrix of mean hitting times by a single number. Note however that $\sum_i \pi_i z_{ij} = 0 \ \forall j$ and so

$$\mathbb{E}_\pi T_j^{\text{hit}} := \sum_i \pi_i \mathbb{E}_i T_j^{\text{hit}} = z_{jj}/\pi_j \tag{4}$$

which in general **does** depend on j.

In the FMIE setting $\mathcal{N}$ is a symmetric matrix, which makes $\mathbf{Z}$ a symmetric matrix, as well as making π be the uniform distribution. But this does not imply that $(\mathbb{E}_i T_j^{\text{hit}})$ is symmetric; in fact

$$\mathbb{E}_i T_j^{\text{hit}} = \mathbb{E}_j T_i^{\text{hit}} \text{ iff } z_{jj} = z_{ii} \text{ iff } \mathbb{E}_\pi T_j^{\text{hit}} = \mathbb{E}_\pi T_i^{\text{hit}}.$$

A chain is **transitive** if for each pair i_1, i_2 there is a permutation σ of the state space such that $\sigma(i_1) = i_2$ and

$$\nu_{ij} = \nu_{\sigma(i),\sigma(j)} \ \forall i, j.$$

Informally, the chain "looks the same from each state". Transitivity implies $z_{ii} = z_{jj} \ \forall i, j$ and hence $T_{ij} \stackrel{d}{=} T_{ji} \ \forall i, j$.
[Board: degrees of freedom].

[RWG Chapter 3]
There are two parallel ways to think about the dynamics of the distribution of $Z(t)$. First, in terms of **matrices**. The transition rate matrix $\mathcal{N}$ has eigenvalues

$$0 = \lambda_1 > -\lambda_2 \geq -\lambda_3 \geq \ldots \geq -\lambda_n$$

and there is a **spectral representation** (matrix diagonalization)

$$\mathbb{P}_i(Z(t) = j) = \pi_i^{-1/2} \pi_j^{1/2} \sum_{m=1}^{n} \exp(-\lambda_m t) u_{im} u_{jm} \tag{5}$$

for orthonormal $\mathbf{U}$. In particular, the time-asymptotics for convergence to stationarity are

$$\mathbb{P}_i(Z(t) = j) - \pi_j = c_{ij} e^{-\lambda_2 t} + o(e^{-\lambda_2 t}) \text{ as } t \to \infty. \tag{6}$$

Jargon: λ_2 is the **spectral gap**, $\tau_{\text{rel}} := 1/\lambda_2$ is the **relaxation time**.

Knowing λ_2 doesn't tell you anything precise about the finite-time distribution of the MC starting at an arbitrary state, but it does tell you some things about the *stationary* chain. For instance (cf. the extremal characterization later)

$$\max_{f,g} \operatorname{cor}_\pi(f(Z(0)), g(Z(t))) = \exp(-\lambda_2 t).$$

Note also

$$\mathbb{P}_i(Z(t) = i) = \pi_i + \sum_{m \geq 2} u_{im}^2 \exp(-\lambda_m t) \tag{7}$$

so the right side is decreasing with t, and in fact is *completely monotone*.

The second way – which we might call "L^2 theory" or "the Dirichlet formalism" – requires some notational setup. For a "test function" g : **Agents** $\to \mathbb{R}$ write

$$\begin{aligned} \bar{g} &= \sum_i \pi_i g_i \\ ||g||_2^2 &= \sum_i \pi_i g_i^2 \\ \mathcal{E}(g,g) &= \tfrac{1}{2} \sum_i \sum_{j \neq i} \pi_i \nu_{ij} (g_j - g_i)^2 \quad \text{(the Dirichlet form).} \end{aligned}$$

When $\bar{g} = 0$ then $||g||_2$ measures "global" variability of g whereas $\mathcal{E}(g,g)$ measures "local" variability relative to the underlying geometry. [Discussion on board]

For a (signed) measure θ we define $||\theta||_{2(m)} = ||f||_2$ for the density $f_i = \theta_i / \pi_i$, and then for a PM μ we have

$$||\mu - \pi||_{2(m)}^2 = -1 + \sum_i \mu_i^2 / \pi_i.$$

This is "L^2 distance" for probability measures.

Why is this viewpoint useful?

The basic evolution equation, for the time-t distribution $\rho(t) = (\rho_j(t))$ from an arbitrary start, is

$$\frac{d}{dt} \rho_j(t) = \sum_j \nu_{ij} \rho_i(t) \tag{8}$$

from which we previously obtained

$$\mathbb{P}_i(Z(t) = j) \text{ are entries of } \exp(\mathcal{N} t).$$

But a little algebra, directly from (8), gives

Lemma (Global convergence equation)

$$\frac{d}{dt} \|\rho(t) - \pi\|_{2(m)}^2 = -2\mathcal{E}(f(t), f(t)); \qquad f_j(t) = \rho_j(t)/\pi_j.$$

Because $\mathcal{E} \geq 0$ this gives a certain "monotonicity" in global convergence; cf. monotonicity of $\mathbb{P}_i(Z(t) = i)$.

Reformulating the classical *Rayleigh–Ritz* extremal characterization of eigenvalues:

Theorem (Extremal characterization of relaxation time)

$$\tau_{rel} = \sup\{||g||_2^2/\mathcal{E}(g,g) : \ \bar{g} = 0\}.$$

So we can get lower bounds on τ_{rel} by plugging in a test function g chosen heuristically. Much of (algorithm-related) uses of finite MCs involves getting reasonable upper bounds on τ_{rel} and τ_{mix} below. The extremal characterization doesn't help directly but is the starting point for other methodologies.
Combining the extremal characterization with the global convergence equation leads easily to [calculation on board]

Lemma (L^2 contraction lemma)

The time-t distributions $\rho(t)$ of a reversible MC satisfy

$$\|\rho(t) - \pi\|_{2(m)} \leq e^{-t/\tau_2}\|\rho(0) - \pi\|_{2(m)}.$$

Hitting times and mixing times are distinct aspects of a MC; but here's a minor connection. Write T_A^{hit} for the hitting time on a subset $A \subset$ **Agents**.

Proposition

For a subset A of a continuous-time chain,

$$\sup_t |\mathbb{P}_\pi(T_A > t) - \exp(-t/\mathbb{E}_\pi T_A)| \leq \tau_{rel}/\mathbb{E}_\pi T_A.$$

In words: for the hitting time distribution to be approximately Exponential it is sufficient that the mean hitting time be large compared to the relaxation time τ_{rel}.

Theory project. Give a bound on the dependence between initial state $X(0)$ and T_A, for instance

$$\max_{f,g} \operatorname{cor}(f(X(0)), h(T_A)) \leq \psi(\tau_{\text{rel}}/\mathbb{E}_\pi T_A).$$

For another connection, recall the *Random Target Lemma* said that $\tau_{\text{hit}} := \sum_j (\mathbb{E}_i T_j^{\text{hit}})\pi_j$ does not depend on i. It turns out that τ_{hit} has a simple expression in terms of the eigenvalues:

$$\tau_{\text{hit}} = \sum^n 1/\lambda_i \quad \textbf{the eigentime identity}. \tag{9}$$

Variation distance (or **total variation**) between a PM μ and the stationary distribution π is defined as

$$||\mu - \pi||_{VD} := \tfrac{1}{2} \sum_i |\mu_i - \pi_i|.$$

This is essentially "L^1 distance". Note that, like "L^2 distance", it ignores the geometry.

For a continuous-time MC, define **(variation distance) mixing time** τ_{mix} to be the smallest time t for which

$$\max_i ||\mathbb{P}_i(Z(t) \in \cdot) - \pi(\cdot)||_{VD} \leq 1/(2e).$$

The choice of constant on the right must be $< 1/2$ but is otherwise rather arbitrary; the particular choice $1/(2e)$ ensures

$$\tau_{\text{rel}} \leq \tau_{\text{mix}}.$$

Variation distance and τ_{mix} are central to many theoretical algorithmic uses of MCs – see Montenegro-Tetali (2006) and the monographs.

The general version of the "bottleneck parameters" earlier are defined in terms of stationary flow rates

$$Q(A, A^c) := \sum_{i \in A, j \in A^c} \pi_i \nu_{ij}.$$

In particular, define the **Cheeger time constant** by

$$\tau_{\text{cond}} := \sup_A \frac{\pi(A)(1 - \pi(A))}{Q(A, A^c)}.$$

[Discussion on board: up to factors of 2 this is 1/conductance; n-cycle case].
There is a (not easy)

Theorem (Cheeger's inequality)

For any continuous-time reversible MC,

$$\tau_{rel} \leq 8\tau_{cond}^2 \max_i \nu_i.$$

So in the FMIE context with the regularity assumption (1) we have

$$\phi(m) = \min\{\nu(A, A^c) : |A| = m\}, \quad 1 \le m \le n-1$$

$$\tau_{\text{cond}} := \sup_m \frac{\frac{m}{n}\frac{n-m}{n}}{\phi(m)}.$$

and Cheeger's inequality becomes

$$\tau_{\text{rel}} \le 8\tau_{\text{cond}}^2.$$

More sophisticated results can be found in the survey by Montenegro-Tetali (2006).

Helpful intuition is that a sequence

$$\hat{Z}_1, \hat{Z}_2, \ldots$$

obtained as either a stationary MC sampled at multiples of τ_{rel}

$$Z(\tau_{\text{rel}}), Z(2\tau_{\text{rel}}), \ldots$$

or an arbitrary-start MC sampled at multiples of τ_{mix}

$$Z(\tau_{\text{mix}}), Z(2\tau_{\text{mix}}), \ldots$$

behaves similarly to an IID sequence as far as quantitative versions of limit theorems are concerned. See e.g. León-Perron (2004) for a large deviation inequality for occupation times.

Another well-studied MC topic is the **cover time**

$$C := \max_j T_j^{\text{hit}} = \text{ time until every state visited.}$$

There is a "non-clever" bound in term of the parameter $\tau^* := \max_{i,j} \mathbb{E}_i T_j^{\text{hit}}$, because inductively

$$\mathbb{P}_i(T_j^{\text{hit}} > 2m\tau^*) \leq 2^{-m}, \ m = 1, 2, 3, \ldots$$

and it quickly follows that

$$\max_i \mathbb{E}_i C \leq (1 + o(1))\tau^* \log n.$$

And a "clever" argument called **Matthews' method** sharpens this to

$$\max_i \mathbb{E}_i C \leq \tau^* \sum_{i=1}^{n-1} 1/i.$$

Recent deep results of Ding-Lee-Peres (2010) give very sharp general estimates of $\mathbb{E}C$.

Reversible Markov chains: standard examples

For any discrete-time MC with transition probabilities p_{ij} there is a corresponding continuous-time MC with transition rates $\nu_{ij} = p_{ij}$. In particular, discrete-time RW on a d-regular undirected graph is the MC with transition probabilities

$$p_{ij} = d^{-1} \text{ for edges } (i, j)$$

and there is a corresponding continuous-time RW.

For a continuous-time MC, in the case where $\nu_i := \sum_{j \neq i} \nu_{ij}$ is constant in i, it is natural to standardize the time unit so that $\nu_i \equiv 1$.

[board: comments re 2 different continuous-times RWs on graphs with highly varying degrees – needs watching in all FMIE contexts]

Continuous-time RW on the complete n-vertex graph.

$$\nu_{ij} = 1/(n-1), \quad j \neq i.$$

The basics are easy – no surprise!

$$\mathbb{E}_i T_j^{\text{hit}} = n-1; \quad T_j^{\text{hit}} \stackrel{d}{=} \text{Exponential}(1/(n-1)).$$

$$\mathbb{P}_i(Z(t) = i) = \tfrac{1}{n} + \left(1 - \tfrac{1}{n}\right) \exp\left(-\tfrac{n}{n-1} t\right).$$

$$\tau_{\text{rel}} = \tfrac{n-1}{n}.$$

Because here $\tau_{\text{rel}} \approx 1$, for other geometries we can think of τ_{rel} as relaxation time relative to the complete graph case.

Continuous-time RW on the d-dimensional lattice/cube/torus.

First consider the infinite lattice $\mathbb{Z}^d$. Discrete-time RW on $\mathbb{Z}^d$ is a well-studied classical object. The continuous-time RW $Z^{(d)}(t)$ is nicer in that the co-ordinate processes are independent slowed-down 1-dimensional RWs; for the origin $\mathbf{0}$ and $\mathbf{x} = (x_1, \ldots, x_d)$

$$\mathbb{P}_{\mathbf{0}}(Z^{(d)}(t) = \mathbf{x}) = \prod_{i=1}^{d} \mathbb{P}_0(Z^{(1)}(t/d) = x_i).$$

Five facts you should know about RW on $\mathbb{Z}^d$.

(CLT): The distribution of $Z^{(d)}(t)$ for large t is approximately multivariate Normal; marginals are Normal$(0, t/d)$.

(Local density) $\mathbb{P}_0(Z^{(1)}(t) = 0) \sim (2\pi)^{-1/2} t^{-1/2}$ and so

$$\mathbb{P}_{\mathbf{0}}(Z^{(d)}(t) = \mathbf{0}) \sim (2\pi t/d)^{-d/2}.$$

(Recurrence/transience) In $d = 1, 2$ RW is **recurrent**: each vertex is visited infinitely often. In $d \geq 3$ RW is **transient**: the chance state $\mathbf{x}$ is ever visited $\to 0$ as $|\mathbf{x}| \to \infty$.

(Fair game: winner and mean duration). In $d = 1$, for $-a < 0 < b$

$$\mathbb{P}_0(T_b < T_{-a}) = a/(a+b); \quad \mathbb{E}_0 \min(T_{-a}, T_b) = ab.$$

Continuous-time RW on the d-dimensional torus.

One natural geometry is the 2-dimensional discrete square $[0, m-1]^2$ as a subgraph of $\mathbb{Z}^2$. It is mathematically nicer to eliminate the boundary by imposing "periodic boundary conditions", that is to use the 2-dimensional discrete torus, which is vertex-transitive. In general dimension $d \geq 1$ this becomes the d-dimensional (discrete) torus, denoted $\mathbb{Z}_m^d$.

Warning; we study $m \to \infty$ asymptotics for fixed d. To compare with other models, remember $n = m^d$.

We quote some basic facts. Consider $d = 1$, so $\mathbb{Z}_m$ is the m-cycle. The eigenvalues are

$$\cos(2\pi j/m), \ 0 \leq j \leq m-1$$

and the relaxation time is

$$\tau_{\text{rel}} = \frac{1}{1 - \cos(2\pi/m)} \sim \frac{m^2}{2\pi^2}.$$

For $d \geq 2$ we retain the nice property that the co-ordinate processes are independent slowed-down versions of the RW on the m-cycle; so for fixed m the d-dimensional and 1-dimensional RWs are again related by

$$\mathbb{P}_{\mathbf{0}}(Z^{(d)}(t) = \mathbf{x}) = \prod_{i=1}^{d} \mathbb{P}_0(Z^{(1)}(t/d) = x_i).$$

From this we see that the eigenvalues on $\mathbb{Z}_m^d$ are

$$\lambda_{(k_1 \ldots k_d)} = \frac{1}{d} \sum_{u=1}^{d} (1 - \cos(2\pi k_u/m)), \ 0 \leq k_u \leq m-1.$$

In particular, the relaxation time satisfies

$$\tau_2 \sim \frac{dm^2}{2\pi^2} = \frac{dn^{2/d}}{2\pi^2}.$$

We can also use the *eigentime identity* to compute the mean hitting time parameter

$$\tau_{\text{hit}} = \sum_{k_1} \cdots \sum_{k_d} 1/\lambda_{(k_1, \ldots, k_d)}$$

(the sum excluding $(0, \ldots, 0)$),

and hence

$$\tau_{\text{hit}} \sim m^d R_d \tag{10}$$

where

$$R_d \equiv \int_0^1 \cdots \int_0^1 \frac{1}{\frac{1}{d} \sum_{u=1}^{d} (1 - \cos(2\pi x_u))} \, dx_1 \cdots dx_d \tag{11}$$

provided the integral converges. In fact by the recurrence/transience properties of RW on the whole integer lattice we must have $R_d < \infty$ for $d \geq 3$ only. For $d = 1$ we must have $\tau_{\text{hit}} = \Theta(m^2)$, and the case $d = 2$ is best understood via a later argument.

Random graphs with prescribed degree distributions

[on board: only key points here]

- Maybe 500 papers since 2000 on such random graph models.
- Configuration model: basic properties and local Galton-Watson approximation.
- Continuous-time vs discrete-time RW
- Mean hitting times via tree recursions. In particular, on random r-regular graph $\tau_{\text{hit}} \sim \frac{r-1}{r-2} n$

Outline. Specify (d_i), Can define models $\mathcal{G}_n$ of n-vertex graph, interpretable as being "random" subject to the following constraint. Write D_n for degree of a uniform random vertex of $\mathcal{G}_n$, then

$$D_n \stackrel{d}{\to} D \text{ where } \mathbb{P}(D = i) = d_i.$$

Such models have the following "local GWBP approximation". The structure of $\mathcal{G}_n$ within some fixed graph-distance r from a uniform random vertex U_n converges in distribution, as $n \to \infty$, to the random tree comprising generations 0 to r of the following modified Galton-Watson BP. The root has offspring distribution D; in subsequent generation the offspring distribution is the size-biased distribution D^* where $\mathbb{P}(D^* = i) = (i+1)d_{i+1}/\mathbb{E}D$.
Assuming $d_0 = d_1 = 0$, the GWBP is an infinite tree (non-extinction).
Assuming $\mathbb{E}D^{2+\varepsilon} < \infty$ then $\mathbb{E}(D^*)^{1+\varepsilon} < \infty$ and the Kesten-Stigum theorem says that the size Y_r of generation r grows at a particular rate: $Y_r/(\mathbb{E}D^*)^r \to W$ a.s. and L^1.

The results above suggest **heuristics** for the structure of $\mathcal{G}_n$ and the behavior of RW and other FMIE processes on $\mathcal{G}_n$.

Let us record the following **local transience principle.** For a large finite-state MC whose behavior near a state i can be approximated be a transient infinite-state chain,

$$\mathbb{E}_\pi T_i^{\text{hit}} \approx R_i/\pi_i$$

where R_i is defined in terms of the approximating infinite-state chain as $\int_0^\infty p_{ii}(t)\ dt = \frac{1}{\nu_i q_i}$, where q_i is the chance the infinite-state chain started at i will never return to i.

The approximation comes from (4) via a "interchange of limits" procedure which requires ad hoc justification.

In the case of simple RW on the $d \geq 3$-dimensional torus, this identifies the constant R_d at (11) as $R_d = 1/q_d$ where q_d is the chance that RW never returns to the origin. So (11) provides a formula for q_d.

In the "random graphs with prescribed degree distribution" model, this argument shows (heuristics) that $\mathbb{E}_\pi T_i^{\text{hit}} = \Theta(n)$.

Other geometries

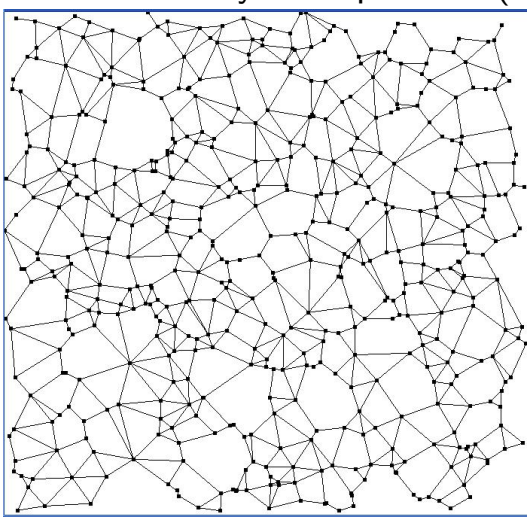
The d-**dimensional hypercube** $\{0,1\}^d$ is often used as the simplest non-trivial example of a geometry on which the RW is rapidly mixing. In particular it illustrates the cut-off window for variation distance mixing [xxx explain on board]. But it seems not so natural for the applications we have in mind.

Small world graphs, which start with the d-dimensional lattice and add random long edges (v, w) with probabilities $\propto ||v - w||^{-\gamma}$, are interesting but hard to study analytically.

Proximity graphs, described next, are also interesting but hard to study analytically.

Given points (x_i) in the plane in general position, create edges according to a deterministic rule such as

create an edge (x_i, x_j) iff the disc $A(x_i, x_j)$ with diameter-line (x_i, x_j) does not contain any third point of (x_i).



Replacing the disc with a one-parameter family of other shapes, and applying this construction to random (Poisson) points, gives a family of **random proximity graphs** which (unlike the more familiar **random geometric graphs**) are always connected.

Simulation project. Study RW and other FMIE processes on these graphs.