Mixing Times and Hitting Times

David Aldous

January 12, 2010

Old Results and Old Puzzles

Levin-Peres-Wilmer give some history of the emergence of this *Mixing Times* topic in the early 1980s, and we nowadays talk about a particular handful of motivations for studying the subject. I'm going to start this talk by relating how I got into this topic. The questions now seem somewhat peripheral to the main themes of the topic, but perhaps should not be completely forgotten.

Donnelly, Kevin P. (1983) *The probability that related individuals share some section of genome identical by descent.*

Let $\{X(l), l \in L\}$ be a random walk indexed by a finite set $L = \{l_i\}$, $l \geq 0$, $1 \leq i \leq 23$, of line segments. The biological significance of l_i is that they are fixed (map) lengths of (human) chromosomes: a chromosome C_i, $1 \leq i \leq 23$, is thought of as a line segment $[0, l_i]$ which may be broken and exchanged by crossover. If the parent chromosomes are labeled by 0 (female) and 1 (male), then the crossover process (or the gene flow) can be described as a continuous parameter random walk on the vertices of a hypercube $H = \{0,1\}^{\mathcal{C}^+}$, where $\mathcal{C}^+$ is the chromosome pedigree structure without the subset $\mathcal{F}$ of founders.

Let F be a mapping $F\colon H \times \mathcal{C} \to \mathcal{F}$ (with the meaning "founder being copied from") which allows one to define the "identity by descent". Under some special biological assumptions, one can say that two chromosomes C_1 and C_2 are "detectably related" if and only if by "time" l_i the random walk X has hit the set $G = \{h \in H\colon F(h, C_1) = F(h, C_2)\}$. The mathematical problem is the computation of the hitting probabilities (G is assumed to be an absorbing set) and the distribution of the absorption time. The inherent difficulties are reduced by introducing a group of symmetries of H which induces a partition of the set of vertices into a set of orbits. Computations are done for the obvious pedigree relationships (e.g., grandparent-, half-sib-, cousin-type, etc.). The probability that an individual with n children passed on all his (her) genes to them equals 0.9 for $n = 13$ and 0.03 for $n = 7$

The math structure in this particular problem is:
Continuous-time random walk on n-dimensional hypercube $\{0,1\}^n$.
Given a subset of vertices $A \subset \{0,1\}^n$, want to study distribution of hitting time τ_A from uniform start.
In this particular problem n is small, ($n =$ "distance in family tree" $= 4$ for cousins) A is small and has symmetry, (e.g. $A = \{0000,\ 1010\}$) and we can get explicit answers.

Back in 1981, this prompted me to ponder

- Instead of some special chain think of a general chain: finite state space Ω, stationary distribution π. To avoid periodicity issues work in continuous-time.
- In principle can calculate dist. of τ_A exactly. But except in special cases, answer (formula for generating function in terms of matrix inverses) not human-readable.
- What can we do instead of exact formulas?

Somehow ...... I started the following line of thought.

Consider mean hitting times $\mathbb{E}_i \tau_A$ as a function of starting state i. This function has "average" $\mathbb{E}_\pi \tau_A = \sum_i \pi_i \, \mathbb{E}_i \tau_A$ which by analogy with IID sampling we expect to be of order $1/\pi(A)$ in the case $\pi(A)$ is small. This "$\pi(A)$ small" case seems the only case we can hope for general results.

Suppose for each initial state i we can find a stopping time $T_{i,\pi}$ at which the chain has distribution π, and set

$$\mathbf{t}_* = \max_i \mathbb{E}_i T_{i,\pi}$$

[On the hypercube we can actually calculate $\mathbf{t}_*$ which reduces to study of the 1-dimensional birth-and-death Ehrenfest urn chain; it is order $n \log n$.] Observe upper bound

$$\mathbb{E}_i \tau_A \le \mathbf{t}_* + \mathbb{E}_\pi \tau_A .$$

General principle (function or RV): if maximum only just larger than the average, then most values are close to average; one implementation give

$$\sum_i \pi_i \left| \frac{\mathbb{E}_i \tau_A}{\mathbb{E}_\pi \tau_A} - 1 \right| \le 2 \frac{\mathbf{t}_*}{\mathbb{E}_\pi \tau_A} .$$

This result is not so impressive in itself, but is perhaps the first (most basic) result using what we now call *a mixing time*, in this case $\mathbf{t}_*$.

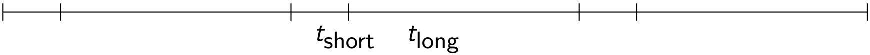
Repeating in words:
for any chain, for any subset A with $\pi(A) \ll 1/\mathbf{t}_*$,

$$\frac{\mathbb{E}_i \ \tau_A}{\mathbb{E}_\pi \tau_A} \approx 1 \text{ for most } i.$$

There's a more natural question in this setting. By analogy with IID sampling we expect the *distribution* of τ_A to be approximately exponential; though local dependence will typically change the mean away from $1/\pi(A)$. Slightly more precisely, we expect:

if $\mathbb{E}_\pi \tau_A \gg$ (a suitable mixing time) then the distribution (starting from π) of τ_A should be approximately exponential with its true mean $\mathbb{E}_\pi \tau_A$.

[Note: then, by previous argument, true for most initial i.]
Here's a brief outline of an argument.
Take $t_{\text{short}} \ll t_{\text{long}} \ll \mathbb{E}_\pi \tau_A$ and divide time into short and long blocks.



Enough to show, from any state i at start of short block,

$$\mathbb{P}_i(\text{visit } A \text{ during next long block}) \approx c \ \ \forall i.$$

[Note: chance τ_A is in some short block $\approx t_{\text{short}}/(t_{\text{long}} + t_{\text{short}})$, by stationarity, so neglect this possibility.]

Write $c = \mathbb{P}_\pi(\tau_A \le t_{\text{long}})$ and note left side $= \mathbb{P}_{\rho(i,t_{\text{short}})}(\tau_A \le t_{\text{long}})$ where $\rho(i, t_{\text{short}})$ is time-t_{short} distribution of chain started at i.

So what we need to make the argument work is that total variation distance $||\rho(i, t_{\text{short}}) - \pi||_{TV}$ is small $\forall i$.

This (and many similar subsequent arguments) motivated definition of "total variation mixing time" $\mathbf{t}_{\text{mix}}$. The argument relies on $\mathbf{t}_{\text{mix}} \ll t_{\text{short}} \ll t_{\text{long}} \ll \mathbb{E}_\pi \tau_A$ and leads to a theorem of the form

$$\sup_t |\mathbb{P}_\pi(\tau_A > t) - \exp(-t/\mathbb{E}_\pi \tau_A)| \le \psi(\mathbf{t}_{\text{mix}}/\mathbb{E}_\pi \tau_A)$$

for a universal function $\psi(\delta) \to 0$ as $\delta \to 0$.

$$(*) \quad \sup_t |\mathbb{P}_\pi(\tau_A > t) - \exp(-t/\mathbb{E}_\pi \tau_A)| \le \psi(\mathbf{t}_{\mathrm{mix}}/\mathbb{E}_\pi \tau_A)$$

for a universal function $\psi(\delta) \to 0$ as $\delta \to 0$.

I once published a crude version of this via method above, but ......

Open Problem 1: Prove a clean version of (*).
You may change left side to some other measure of distance between distributions; you may change definition of mixing time; but want optimal order of magnitude for $\psi(\delta)$.

Turn to **reversible** chains. Recall the notion of **spectral gap** λ; characterized e.g. via "maximum correlation for the stationary chain":

$$\max_{f,g} \mathrm{cor}_\pi(f(X_0), g(X_t)) = \exp(-\lambda t).$$

λ has dimensions "1/time" ; to get a quantity with dimensions "time" set

$$\mathbf{t}_{\mathrm{rel}} = 1/\lambda = \text{"relaxation time"}.$$

For reversible chains there is a remarkable clean version of (*)

$$\sup_t |\mathbb{P}_\pi(\tau_A > t) - \exp(-t/\mathbb{E}_\pi \tau_A)| \le \mathbf{t}_{\mathrm{rel}}/\mathbb{E}_\pi \tau_A$$

For a reversible chain with relaxation time $\mathbf{t}_{\mathrm{rel}}$

$$(**) \quad \sup_t |\mathbb{P}_\pi(\tau_A > t) - \exp(-t/\mathbb{E}_\pi \tau_A)| \leq \mathbf{t}_{\mathrm{rel}}/\mathbb{E}_\pi \tau_A$$

This has a “non-probabilistic” proof (see Aldous-Fill Chapter 3).

Open (no-one has thought about . . .) **Problem 2:**

In the setting of (**) give a bound on the dependence between initial state X_0 and τ_A, for instance

$$\max_{f,h} \mathrm{cor}_\pi(f(X_0), h(\tau_A)) \leq \psi(\mathbf{t}_{\mathrm{rel}}/\mathbb{E}_\pi \tau_A).$$

Some “probabilistic” proof of (**) might also answer this.

I have shown 3 results using 3 different formalizations of *mixing time*. Back in 1981 this was a bit worrying, so I put a lot of effort into thinking about variants and their relationships.

- $\mathbf{t}_{\mathrm{mix}}$ involves choice of "variation distance" as well as arbitrary numerical cutoff.
- $\mathbf{t}_* = \max_i \mathbb{E}_i T_{i,\pi}$ is less arbitrary, but harder to use.

Another view of the concept of a mixing time $\mathbf{t}$:
"sampling a chain at time-intervals $\mathbf{t}$ should be as good as getting IID samples at these intervals".
Different choices of what you're wanting to do with the samples lead to different definitions of $\mathbf{t}$; considering mean hitting times leads to the definition

- $\mathbf{t}_{\mathrm{hit}} = \max_{i,A} \pi(A)\ \mathbb{E}_i \tau_A$

Aldous (1982) shows: for continuous-time **reversible** chains, these 3 numbers $\mathbf{t}_{\mathrm{mix}}, \ \mathbf{t}_*, \ \mathbf{t}_{\mathrm{hit}}$ are equivalent up to multiplicative constants.

Lovasz-Winkler (1995) studied other mixing times based on hitting times. In the non-reversible case, there are 2 families of internally-equivalent parameters, and the families "switch" under time-reversal.

Bottom line:
Nowadays we have settled on a definition of "total variation mixing time" $\mathbf{t}_{\text{mix}}$ as smallest time t for which

$$\max_i ||P_i(X_t \in \cdot) - \pi(\cdot)||_{TV} \leq 1/(2e)$$

(or $\leq 1/4$ in discrete time, more commonly). The type of results just mentioned (equivalence up to constants) provide some justification for "naturalness", though

Puzzle 3: the actual results seem rarely useful in bounding $\mathbf{t}_{\text{mix}}$.

For instance one can choose an arbitrary distribution ρ instead of π and bound $\mathbf{t}_{\text{mix}}$ via constant times $\max_i \mathbb{E}_i T_{i,\rho}$. (This is closely analogous to the standard treatment of recurrence for general-space chains.) But there are very few examples where this method is useful.

Anyway (*), there is a unique "order of magnitude" for $\mathbf{t}_{\text{mix}}$ for families parametrized by size, for instance order $n \log n$ for the n-dimensional hypercube $\{0,1\}^n$, and the typical modern use of mixing times is, within a family parametrized by size, to use known order of magnitude of mixing times to help estimate order of magnitude of some other quantity of interest.

The relaxation time $\mathbf{t}_{\mathrm{rel}}$ is usually different – order n in the hypercube case – though (as here) usually not *much* different from $\mathbf{t}_{\mathrm{mix}}$.

Very vaguely, the "equivalence" theory for $\mathbf{t}_{\mathrm{mix}}$ is working in the L^1 and L^∞ worlds – look at $\max_i ||P_i(X_t \in \cdot) - \pi(\cdot)||_{TV}$, whereas the relaxation time $\mathbf{t}_{\mathrm{rel}}$ is working in L^2 theory – look at

$$\max_{f,g} \mathrm{cor}_\pi(f(X_0), g(X_t)) = \exp(-t/\mathbf{t}_{\mathrm{rel}}).$$

Puzzle 4: Why isn't there a parallel L^2 theory, for reversible chains at least, relating the relaxation time $\mathbf{t}_{\mathrm{rel}}$ to non-asymptotic L^2 properties of hitting times?

There are many equivalent characterizations of $\mathbf{t}_{\mathrm{rel}}$, but not in terms of hitting times.

Hitting times is itself just a small topic within *Markov chains*, but it does relate to some other topics.

Coalescing random walks.

Reversible continuous-time Markov chain with finite state space.
Start one particle from each state; particles coalesce if they meet.
Study random time C at which all particles have coalesced into one.

Model interesting for two reasons:

- Dual to voter model
- Kingman's coalescent is "complete graph" case.

Parameter $\mu =$ mean time for two π-randomly started particles to meet.

Open Problem 5: Suppose $\mu \gg \mathbf{t}_{\mathsf{rel}}$. Under what extra assumptions can we show $\mathbb{E}C \approx 2\mu$?

We expect this because

- Meeting time of 2 particles has approx. Exponential (mean μ) distribution
- with k particles, there are $\binom{k}{2}$ pairs, each pair meets at Exponential (mean μ) random time, so if these times were independent then the first such meeting times would have approx. Exponential (mean $\mu/\binom{k}{2}$) dist.
- $\sum_{k\geq 2} 1/\binom{k}{2} = 2$

and proved by Cox (1989) on torus $\mathbb{Z}_n^d$, ($d \geq 2$ fixed) using more explicit calculations.

Previous results/problems relevant to understanding what happens starting with fixed k, w.l.o.g. $k = 3$ (different argument needed to show contribution from large k is negligible).

Meeting time of 2 particles is a hitting time for the bivariate process (X^1_t, X^2_t), which is reversible with the same relaxation time $\mathbf{t}_{\mathrm{rel}}$, so from stationary $\pi \times \pi$ start we do have the Exponential (mean μ) approximation.

For 3 particles, first meeting time of some 2 particles is a hitting time for the trivariate process (X^1_t, X^2_t, X^3_t), which is reversible with the same relaxation time $\mathbf{t}_{\mathrm{rel}}$, so from stationary $\pi \times \pi \times \pi$ start we do have the Exponential (mean $=$????) approximation.

But ......why is mean $\approx \mu/3$ – do we need to go via proving some explicit approximate independence property for the 3 meeting times (Open Problem 2) or is there a direct way?

And ......after the first coalescence, need some “approximate $\pi \times \pi$” for the distribution of the 2 particles, to use the 2-particle result.

Is there is an elegant argument using $\mathbf{t}_{\mathrm{rel}}$? If we work with $\mathbf{t}_{\mathrm{mix}}$ we can just combine the ideas above with the crude short block/long block argument.

Cerny - Gayrard (2008) *Hitting time of large subsets of the hypercube.*

Summary: “We study the simple random walk on the n-dimensional hypercube, in particular its hitting times of large (possibly random) sets. We give simple conditions on these sets ensuring that the properly rescaled hitting time is asymptotically exponentially distributed, uniformly in the starting position of the walk. These conditions are then verified for percolation clouds with densities that are much smaller than $(n \log n)^{-1}$. A main motivation behind this article is the study of the so-called aging phenomenon in the Random Energy Model (REM), the simplest model of a mean-field spin glass. Our results allow us to prove aging in the REM for all temperatures.