Markov chains and mixing times

Vaughn Climenhaga

February 15, 2013

Our seminar series is taking a hiatus from spectral methods for a couple weeks – these will return eventually, but in the meantime we'll spend some time with the idea of *coupling* as a method for deriving statistical properties. In this week's post, based on Matt Nicol's talk, we discuss Markov chains and the idea of mixing times.

1 Markov chains and probability distributions

For our purposes, a Markov chain is a (finite or countable) collection of states S and transition probabilities p_{ij}, where $i, j \in S$. We write $P = [p_{ij}]$ for the matrix of transition probabilities. Elements of S can be interpreted as various possible states of whatever system we are interested in studying, and p_{ij} represents the probability that the system is in state j at time $n+1$, if it is state i at time n. We will think of a Markov chain as a stochastic process with state space $S^{\mathbb{N}}$, representing all sequences $X_0, X_1, X_2, \ldots$, where X_n is the state of the system at time n. The characterisation just given of the transition probabilities can be expressed as

$$P(X_{n+1} = j \mid X_n = i) = p_{ij}, \tag{1}$$

where $P(\cdot \mid \cdot)$ represents conditional probability. It is a key feature of a Markov chain (as opposed to other kinds of stochastic processes) that once X_n is known, X_{n+1} is completely independent of any information on what happened before time n – that is, conditioning the probability in (1) on any events involving $X_0, \ldots, X_{n-1}$ does not change its value.

It is tempting to phrase the above property as "X_{n+1} depends only on X_n". However, this formulation is a little misleading, as it gives the impression that the sequence $X_n \in S$ evolves deterministically, whereas it is of

course a stochastic process. To make a correct statement along these lines, we must say that "the probability distribution of X_{n+1} depends only on the probability distribution of X_n".

Let us expand this idea. If we write $p_n(j)$ for the probability that the system is in state j at time n, then $p_n \colon S \to [0,1]$ has the property that $\sum_{j \in S} p_n(j) = 1$. That is, p_n is an element of the unit simplex in $\mathbb{R}^d$ (if S has d elements) or in ℓ^1 (if S is countably infinite). Write Δ for this unit simplex: then the sequence of probability distributions p_n can be viewed as a sequence of points in Δ.

Now the Markov property – the fact that the probability in (1) does not change if we condition on X_k for $k < n$ – can be reformulated as the property that $p_{n+1} \in \Delta$ depends only on $p_n \in \Delta$, and not on p_k for any $k < n$. In particular, the Markov chain can be viewed as a (deterministic!) map $T \colon \Delta \to \Delta$, so once we specify the initial probability distribution p_0, subsequent distributions are determined by $p_n = T^n p_0$. Our goal will be to understand how the distributions p_n evolve in time – that is, what are the dynamics of the map T.

The map T is determined by the transition probabilities p_{ij} (and vice versa). In fact, it can be written in terms of the matrix P using the fact that

$$\begin{aligned} p_{n+1}(j) &= P(X_{n+1} = j) \\ &= \sum_{i \in S} P(X_{n+1} = j \mid X_n = i) \cdot P(X_n = i) \\ &= \sum_{i \in S} p_{ij} p_n(i), \end{aligned}$$

so that using the language of matrix multiplication we have

$$p_{n+1} = p_n P, \qquad T(\mu) = \mu P, \tag{2}$$

where μ represents an arbitrary distribution in Δ. We conclude that iterates of the map T correspond to powers of the matrix P.

2 Stationary distribution(s)

It is natural to ask if there is a *stationary distribution* – that is, a distribution $\mu \in \Delta$ such that if we start the Markov chain in this distribution and then run it forward in time, we keep the same distribution, so that $P(X_n = j) =$

$P(X_0 = j)$ for every $j \in S$ and $n \in \mathbb{N}$. If the chain is in such a distribution, then we have a prediction of the long-term behaviour – at any time n, no matter how far in the future, we know how likely it is to find the system in a given state.

A related question is to ask about the long-term behaviour when the initial distribution is *not* stationary. If we start the chain in a distribution p_0 such that $p_1 \neq p_0$, and the probabilities really are changing in time, what happens to $p_n(j)$ as n grows? Does the probability of finding the system in state j at some later time n depend critically on just when we decide to make our observation? Or does $p_n(j)$ converge to some asymptotic probability?

We will answer the first question (existence of stationary distributions) in this section, and address the second later on.

A distribution $\mu \in \Delta$ is stationary for the Markov chain if and only if $T(\mu) = \mu$ – that is, μ is a fixed point for T. In any convex space (such as Δ), there is a method for looking for fixed points: begin with *any* distribution μ, and then consider the Cèsaro averages

$$\mu_n = \frac{1}{n} \sum_{k=0}^{n-1} T^k(\mu). \tag{3}$$

It is not difficult to show that any limit point of this sequence is a fixed point for T – that is, $T(\pi) = \pi$ if $\pi = \lim_{\ell \to \infty} \mu_{n_\ell}$ for some $n_\ell \to \infty$.

To prove this one first needs to clarify what notion of convergence is being used. Consider for now the case when S is finite, say $\#S = d$. In this case $\mu_n \in \mathbb{R}^d$ and we can use the usual topology from $\mathbb{R}^d$, so that in particular the simplex Δ is compact and the sequence μ_n has a convergent subsequence μ_{n_ℓ}. It is a worthwhile exercise to check that the limit $\pi = \lim_{\ell \to \infty} \mu_{n_\ell}$ is in fact a fixed point, $\pi = \pi P$.

Thus we have existence of a stationary distribution when S is finite. There are several natural questions to ask at this point, and we address each in turn.

1. Is the stationary distribution unique? Or can there be multiple stationary distributions?

2. What happens when S is countably infinite? Is there a stationary distribution? Is it unique?

3. If the stationary distribution π is unique, then the averaged distributions μ_n from (3) converge to π. What about the distributions $T^k(\mu)$

themselves, which are just the observations made at time k, without averaging over k? Do these converge to π?

The answers to these questions are as follows: definitions of the relevant terms will come in a moment.

1. The stationary distribution is unique if the Markov chain is *irreducible*.
2. If S is countably infinite then there is a stationary distribution if and only if the Markov chain is *positive recurrent*. If in addition the Markov chain is irreducible then this stationary distribution is unique. For *null recurrent* and *transient* chains, there is no stationary distribution.
3. If the Markov chain is *aperiodic*, then the distributions $T^k(\mu)$ converge to π for every initial distribution μ.

3 Irreducibility

One can associate a directed graph $\mathcal{G}$ to a Markov chain by taking S for the set of vertices and drawing an edge from i to j if and only if the transition probability p_{ij} is non-zero.

Now the Markov chain can be interpreted as a random walk on the graph $\mathcal{G}$. If at time n the random walker is at the vertex labeled i, then the probability that he walks to vertex j at time $n+1$ is given by the transition probability p_{ij}.

Definition 1 *The Markov chain is irreducible if its associated graph $\mathcal{G}$ is strongly connected – that is, there is a path from any vertex to any other vertex.*

The path in Definition 1 may be of any length: an equivalent formulation is that the Markov chain is irreducible if and only if for every $i, j \in S$ there exists $n \in \mathbb{N}$ such that $P(X_n = j \mid X_0 = i) > 0$.

Theorem 2 *If S is finite and the Markov chain is irreducible, then there is a **unique** stationary distribution $\pi = \pi P$ on S.*

Theorem 2 is part of the Perron–Frobenius theorem in linear algebra.

4 An example – card shuffling

Consider a deck of n cards, and let S be the set of all possible orderings of those cards, so that $\#S = n!$. The act of shuffling the deck can be described as a Markov chain with state space S: if $i \in S$ represents the current order of the cards, then the acts of shuffling once changes the order to some other element of S, and the probability of transitioning from the ordering $i \in S$ to the ordering $j \in S$ encodes some properties of the method of shuffling employed.

One of the simplest possible shuffles is the *top-down shuffle* – given a deck in state i, choose a card at random, remove it from the deck, and place it on the top. Thus there are n possible states that can be reached in a single step from the current state, and each one is equally likely. In terms of the random walk on a graph described in the previous section, we have a directed graph $\mathcal{G}$ with $n!$ vertices, each with outgoing degree n, and the random walker selects one of the n outgoing edges uniformly at random at each step.

It is easy to see that this Markov chain is irreducible – given any two states it is possible to go from one to the other (in enough steps) with positive probability. So there is a unique stationary distribution. What is it? Intuitively we feel as though the shuffling process is symmetric enough (blind enough to the details of the arrangement at any given time) that all arrangements of the cards should be equally likely. In other words, we expect that the stationary distribution is the $\pi \in \Delta$ defined by $\pi(i) = \pi(j)$ for all $i, j \in S$ – that is, $\pi(j) = \frac{1}{n!}$ for all j.

To see that this distribution is stationary, define for each $j \in S$ the set $I(j) = \{i \in S \mid i \to j \text{ is an edge in } \mathcal{G}\}$ of configurations of the deck from which the configuration j can be reached with a single step of the shuffle (a single selection of a card). Observe that $I(j)$ has exactly n elements, corresponding to the n positions in the deck from which the top card in configuration j could have been prior to the last shuffle. Thus every vertex in the graph $\mathcal{G}$ has the same incoming degree n, and we can compute πP as

$$(\pi P)(j) = \sum_{i \in I(j)} \pi_i p_{ij} = \frac{1}{n!} = \pi(j).$$

Thus π is indeed the unique stationary distribution.

Remark 3 *The result that every vertex has the same incoming degree can*

also be derived more abstractly as a consequence of the fact that the symmetric group on n elements acts transitively on the graph $\mathcal{G}$.

It is natural to ask how many times we must shuffle in order to feel confident that the resulting distribution is "random enough". In terms of the random walk on the graph $\mathcal{G}$, the act of shuffling looks like this: we start with a probability distribution concentrated on a single vertex $j \in S$, corresponding to the initial ordering of the cards. After a single shuffle, that distribution is evenly distributed over the n vertices that can be reached from j in a single step. After a second shuffle, it is evenly distributed over the n^2 vertices that can be reached from j in two steps – except that it is not quite an even distribution now, because some vertices can be reached in multiple ways via a path of length two. For example, j itself can be reached in two ways (select the top card twice, or select the second card from the top twice).

As the number of shuffles increases, the distribution gets spread out over more and more vertices, but there is also this phenomenon of recurrence where it comes back to certain vertices more quickly than to others, and does not spread out completely evenly. We would like to understand if it eventually approaches the stationary distribution π, which is the uniform distribution on S, and if so, how quickly it does so. We will return to this below.

5 Recurrence and transience

When the set of states S is infinite, the simplex $\Delta \subset \ell^1$ is no longer compact, and so the sequence μ_n in (3) does not necessarily have a convergent subsequence. In particular, the proof of existence for a stationary distribution given above does not immediately go through.

Indeed, consider the Markov chain with state space $S = \mathbb{N}$ and transition probabilities $p_{ij} = 1(j = i + 1), 0$ (otherwise). Then the walk is not so random – the walker simply goes from state i to state $i+1$ at each time step, and in particular we have $P(X_n \leq n) = 0$, so $\lim_{n\to\infty} \mu_n(j) = 0$ for every $j \in S$, and we see that the distributions μ_n converge pointwise to the zero distribution. This is not a probability distribution (the total mass is 0, not 1), and informally we may say that the missing mass has escaped to infinity.

So we need something to replace compactness of Δ, something that will guarantee that no mass escapes to infinity. We may think of this in the

language of the previous post – how do we gaurantee precompactness of the sequence μ_n? What conditions on a subset of ℓ^1 guarantee that it is precompact?

First we note that because we are now in ℓ^1, not $\mathbb{R}^d$, the notion of convergence and of metric have changed. The relevant metric in this case is the ℓ^1 metric

$$d_{\ell^1}(\nu, \mu) = \|\nu - \mu\|_{\ell^1} = \sum_{j \in S} |\nu(j) - \mu(j)|. \tag{4}$$

It is helpful to interpret this metric in light of the fact that $\nu, \mu \in \Delta$ are to be thought of as probability distributions. Upon restriction to Δ, the metric (4) (or rather, $\frac{1}{2}$ this metric) becomes the total variation metric on the space of probability measures:

$$d_V(\nu, \mu) = \frac{1}{2} d_{\ell^1}(\nu, \mu) = \sup_{A \subset S} |\nu(A) - \mu(A)| = \sup_{A \subset S} (\nu(A) - \mu(A)), \tag{5}$$

where we will prove momentarily that the last three expressions are equivalent. First note that $0 \leq d_V(\nu, \mu) \leq 1$ for all $\nu, \mu \in S$. Moreover, if X, Y are random variables on S distributed according to ν, μ respectively, then

$$d(\nu, \mu) = \sup_{A \subset S} |P(X \in A) - P(Y \in A)|,$$

and once again the supremum is unchanged if we remove the absolute value signs, as we now see.

Proposition 4 *The quantities in (5) all coincide.*

Proof: To see that the last two coincide, we observe that if $A^c = S \setminus A$ is the complement of A, then because μ, ν are probability distributions we have

$$\nu(A^c) - \mu(A^c) = (1 - \nu(A)) - (1 - \mu(A)) = \mu(A) - \nu(A).$$

Thus the set over which the last supremum is taken is symmetric around 0.

To see that this supremum coincides with $\frac{1}{2} d_{\ell^1}(\nu, \mu)$, we observe that

$$\nu(A) - \mu(A) = \sum_{i \in S} \mathbf{1}_A(i)[\nu(i) - \mu(i)],$$

and the right-hand side is maximised when $A = \{i \in S \mid \nu(i) \geq \mu(i)\}$. Writing $\hat{A}$ for this set, we have

$$\begin{aligned}\sup_{A \subset S}(\nu(A) - \mu(A)) &= \nu(\hat{A}) - \mu(\hat{A}) \\ &= \frac{1}{2}[(\nu(\hat{A}) - \mu(\hat{A})) + (\mu(\hat{A}^c) - \nu(\hat{A}^c))] \\ &= \frac{1}{2}\sum_{i \in S}\left[\mathbf{1}_{\hat{A}}(\nu(i) - \mu(i)) + \mathbf{1}_{\hat{A}^c}(\mu(i) - \nu(i))\right] \\ &= \frac{1}{2}\sum_{i \in S}|\nu(i) - \mu(i)| = \frac{1}{2}d_{\ell^1}(\nu, \mu).\end{aligned}$$

□

If $\mu_n, \mu \in \Delta$ are such that $d_V(\mu_n, \mu) \to 0$ as $n \to \infty$, then we say that μ_n converges to μ in the total variation norm. This implies pointwise convergence ($\mu_n(j) \to \mu(j)$ for every $j \in S$), but is stronger. For example, the sequence of distributions at the beginning of this section converges to 0 pointwise, but not in total variation.

Now that we know what metric we are using, we can return to the question of precompactness. In fact it turns out that the same Kolmogorov–Riesz theorem that was mentioned in the previous post on compactness comes to our rescue here, although we need a different aspect of it. In that setting, we worked with L^p spaces over a probability measure, and found that precompactness of $\mathcal{F} \subset L^p$ was equivalent to boundedness and uniformly small variation under small perturbations of the argument. In our setting here, boundedness is guaranteed by the fact that every $\mu \in \Delta \subset \ell^1$ has $\|\mu\|_{\ell^1} = 1$, and interpreting sequences in ℓ^1 as functions $\mathbb{N} \to \mathbb{R}$, we see that the argument is discrete, and so that there are no small perturbations to worry about.

However, in our case the underlying measure of the L^p space (ℓ^1) is the counting measure on the integers, which is not a probability measure, but rather is σ-finite. For such measures there is an extra condition in the Kolmogorov–Riesz compactness theorem, which for ℓ^1 can be stated as follows.

Theorem 5 *A subset $\mathcal{F} \subset \Delta \subset \ell^1$ is precompact if and only if for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $\sum_{j \geq N} \mu(j) < \epsilon$ for all $\mu \in \mathcal{F}$.*

A set $\mathcal{F}$ satisfying the hypothesis of Theorem 5 is called tight. If X_μ is a random variable on S distributed according to μ, then tightness can be reformulated as the requirement that for every $\epsilon > 0$, there exists a finite subset $K \subset S$ such that $P(X_\mu \notin K) < \epsilon$ for all $\mu \in \mathcal{F}$. This is the precise condition that keeps the sequence of measures from "escaping to infinity".

Remark 6 *One can formulate the condition of tightness for measures on any topological space, by demanding that the subset K be compact (which is equivalent to finite when the space has the discrete topology, as with our state space S). For measures on a separable metric space, Prokhorov's theorem states that tightness is equivalent to precompactness in the weak* topology. In general this is weaker than precompactness in the total variation norm, because a sequence of measures can converge in the weak* topology without converging in total variation. However, for ℓ^1 the two topologies coincide because the underlying metric space S is discrete.*

How do we verify tightness for the sequence of measures μ_n in (3)? Although we shall not describe the entire theory here, it is worth at least mentioning some of the relevant terminology.

Given a Markov chain with countably infinite state space S, fix a state $j_0 \in S$ and consider the random walk on the associated graph $\mathcal{G}$ that is associated to the Markov chain. Let $R(j_0)$ be a random variable describing the first time at which the walk returns to j_0 – that is,

$$R(j_0) = \min\{n \geq 1 \mid X_n = j_0\},$$

where $X_0 = j_0$ with probability 1. Note that $R(j_0)$ takes the value ∞ if the walk never returns to j_0. Given $n \geq 1$, let $W_n(j_0) = P(R(j_0) > n)$ be the probability that the walker has not returned to state j_0 by time n – this can be computed from the graph $\mathcal{G}$ by taking a sum over all paths of length n that start at the vertex j_0 and do not return to it.

Now one of the following three things happens, and it turns out that which of these three cases happens does not depend on the choice of j_0 (as long as the chain is irreducible):

- $\sum_{n\geq 1} W_n(j_0) < \infty$. In this case the chain is called *positive recurrent*: the return time $R(j_0)$ is finite with probability 1, and the expected value of R is also finite.

- $W_n(j_0) \to 0$ but $\sum_{n\geq 1} W_n(j_0) = \infty$. In this case the chain is called *null recurrent*: the return time is finite with probability 1, but the expected value of R is infinite.

- $W_n(j_0) \not\to 0$. In this case the chain is called *transient*: the return time is infinite with probability 1.

It turns out (though we shall not prove it here) that the chain has a stationary probability distribution π if and only if it is positive recurrent, and that in this case (given irreducibility) π is unique.

6 Aperiodicity and mixing times

From now on we assume that the Markov chain has a unique stationary distribution – that is, it is irreducible and either S is finite or the chain is positive recurrent. Then we have $\mu_n \to \pi$, the unique stationary distribution, for the sequence (3), no matter what the initial distribution μ is. But what happens to $T^n(\mu) = \mu P^n$, the probability distribution of X_n when X_0 is distributed according to μ? Do we really need the averaging process in (3) to get convergence? Or do we get convergence if we take measurements at a single time n?

Definition 7 *A Markov chain is aperiodic if there is no integer > 1 that divides the length of every cycle in the associated directed graph.*

The Perron–Frobenius theorem, mentioned above, states that if π is the unique stationary distribution for an aperiodic Markov chain and $\mu \in \Delta$ is any initial distribution, then $\mu P^n \to \pi$ in total variation norm as $n \to \infty$.

Given $x \in S$, let p_x^m be the distribution that results from starting at state x and running the Markov chain for m steps. Consider the quantities

$$\begin{aligned} D_x(m) &= d_V(p_x^m, \pi), \\ D(m) &= \sup_{x\in S} D_x(m), \\ \tau_x(\epsilon) &= \min\{m \mid D_x(m) \leq \epsilon\}, \\ \tau(\epsilon) &= \sup_{x\in S} \tau_x(\epsilon). \end{aligned}$$

Then $\tau(\epsilon)$ is the minimum time required for every initial condition to lead to a probability distribution that is within ϵ of the stationary distribution

(using the total variation norm). This is the *mixing time* of the Markov chain.

As ϵ decreases, $\tau(\epsilon)$ increases, and we are interested in this rate of growth. If $\tau(\epsilon)$ grows polynomially in $-\log \epsilon$, then the Markov chain is called *rapidly mixing*. We will study this property more in next week's talk. For now we just observe that in the card-shuffling example, this will give us a reasonable measure of how many shuffles it takes for the deck to be "well shuffled".

Note that in that example, the requirement that $d_V(p_x^m, \pi) < \epsilon$ is quite strong. The underlying graph has $n!$ vertices, and the stationary distribution gives each one weight $\frac{1}{n!}$. Thus the initial distribution, which is a delta distribution on a single vertex, must spread out until every vertex has a weight $p_x^m(j) \in [\frac{1}{n!} - \epsilon, \frac{1}{n!} + \epsilon]$.

As an example of how the total variation norm behaves in this case, notice that if we play with a regulation deck of 52 cards and consider a probability distribution μ on the space S of all possible orderings for which a particular card, say the ace of spades, is on the top of the deck with probability 1, then we can estimate the total variation distance from π by using the event $B = \{j \in S \mid \text{ the ace of spades is on top of the deck in the ordering } j\}$, and get

$$d_V(\mu, \pi) \geq \mu(B) - \nu(B) \geq 1 - \frac{51!}{52!} = \frac{51}{52}.$$

Thus μ is almost as far from π as is possible in this metric, and being within a small ϵ of π corresponds to having almost no information about the ordering of the cards.