Markov chains and mixing times (part 2 - coupling)

Vaughn Climenhaga

February 21, 2013

This week's post continues last week's discussion of Markov chains and mixing times, and introduces the idea of coupling as a method for estimating mixing times. We remark that some nice notes on the subject of coupling (and others) can be found on Steve Lalley's web page – of particular relevance for our purposes here are the notes on "Convergence rates of Markov chains". A more thorough and complete reference is the book Markov Chains and Mixing Times by D. Levin, Y. Peres, and E. Wilmer.

1 Markov chains as stochastic processes

In the previous post, we characterised a Markov chain as a pair (S, P), where S is a finite or countable state space and $P \in [0,1]^{S\times S}$ is a matrix whose entries p_{ij} represent the transition probabilities between the various states in S. This then allowed us to interpret the Markov chain as a (deterministic) map $T\colon \Delta \to \Delta$, where Δ is the simplex of probability measures on S, and the map T is given by right multiplication by the matrix P.

Another way to describe a Markov chain is using the language of stochastic processes. A stochastic process is a sequence of random variables $X_n\colon \Omega \to S$, where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Such a process is said to be a Markov chain with state space S and transition matrix P if for any $i_1, \dots, i_{n+1} \in S$, we have

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_k = i_k \text{ for all } 0 \le k \le n) = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n). \tag{1}$$

Note that there is a potential source of confusion here in the terminology – to each pair (S, P) we can associate many distinct stochastic processes satisfying (1). Indeed, there is one such process (up to isomorphism) for every initial probability distribution (the probabilities on X_0). Thus whenever the term "Markov chain" is used, here or in the literature, one should be careful to determine whether or not the object defined refers to a single stochastic process with specified initial distribution, or to a collection of stochastic processes all evolving according to the same transition probabilities.

In the language of the previous post, this distinction takes the following form: the pair (S, P) determines a map $T\colon \Delta \to \Delta$, and the individual stochastic processes just described correspond to fixing an initial distribution $p \in \Delta$ and considering the single trajectory of this map given by $p_n = T^n(p) = pP^n$. Thus "Markov chain" may refer either to a single orbit of this map, or to the space of all orbits of the map.

We will not attempt to introduce new terminology to resolve this ambiguity here. Rather, we shall let the context indicate which interpretation is meant – "the Markov chain X_n" will refer to a single stochastic process (including a fixed initial distribution), while "the Markov chain (S, P)" will refer to the set of all stochastic processes distributed according to (1). We will sometimes say that "X_n is a Markov chain over (S, P)" if it is a stochastic process satisfying (1).

Although it is customary to refer to a random process simply by the random variable X_n, we remark that a key ingredient in the process is the probability distribution $\mathbb{P}$ on the measurable space $(\Omega, \mathcal{F})$. In our setting we can always take $\Omega = S^{\mathbb{N}}$ to be the space of infinite sequences of states in S, and then $X_n\colon S^{\mathbb{N}} \to S$ is simply the map that picks out the nth coordinate of the sequence. We will adopt this convention throughout and will write $\mathbb{P}_X$ for the probability distribution on $S^{\mathbb{N}}$ that is implicit in every mention of the random variable X_n, and we will sometimes refer to $\mathbb{P}_X$ as "a Markov chain over (S, P)".

2 Coupling

Consider a Markov chain (S, P). We want to understand the mixing time of this Markov chain – that is, if π is the stationary distribution and we write p_x^m for the distribution associated to starting in state x (with probability 1) and evolving the Markov chain for m steps, then we want to understand the

function
$$\tau(\epsilon) = \min\{m > 0 \mid d_V(p_x^m, \pi) < \epsilon \text{ for all } x \in S\}, \tag{2}$$
where we recall that d_V is the total variation distance $d_V(\mu, \nu) = \sup_{A \subset S} |\mu(A) - \nu(A)|$.

Roughly speaking, the idea behind coupling is to run two copies of the Markov chain simultaneously in such a way that each copy obeys the original transition probabilities, but the two copies nevertheless "communicate" in such a way that they eventually mirror each other. Then the mixing time $\tau(\epsilon)$ can be estimated in terms of the probability that the two copies take at least time τ before this mirroring begins.

Let us make this more precise. A *coupling* of the Markov chain (S, P) is a Markov chain Z_n with state space $S \times S$ such that writing $Z_n = (X_n, Y_n)$, we have

$$\mathbb{P}(X_{n+1} = j \mid Z_n = (i, i')) = \mathbb{P}(X_{n+1} = j \mid X_n = i) = p_{ij} \tag{3}$$

and similarly

$$\mathbb{P}(Y_{n+1} = j' \mid Z_n = (i, i')) = \mathbb{P}(Y_{n+1} = j' \mid Y_n = i') = p_{i'j'}. \tag{4}$$

That is, both X_n and Y_n are Markov chains over (S, P). If we write Q for the transition matrix on $S \times S$ that governs Z_n, then (3) and (4) can be translated into conditions on the coordinates of Q:

$$\sum_{j' \in S} q_{(i,i'),(j,j')} = p_{ij}, \qquad \sum_{j \in S} q_{(i,i'),(j,j')} = p_{i'j'}. \tag{5}$$

Remark 1 *One can also consider couplings where Z_n is not required to be a Markov chain, but only a stochastic process with state space $S \times S$ whose marginal distributions are Markov chains over (S, P). We will not use such couplings, though.*

Recall that we use the convention that the random variables X_n and Y_n are defined as the nth coordinate projections on $S^{\mathbb{N}}$, and Z_n is defined similarly on $(S \times S)^{\mathbb{N}}$. Thus all of the information about the Markov chains X_n, Y_n, Z_n is really carried by the probability distributions $\mathbb{P}_X$, $\mathbb{P}_Y$ on $S^{\mathbb{N}}$ and $\mathbb{P}_Z$ on $(S \times S)^{\mathbb{N}} = (S^{\mathbb{N}}) \times (S^{\mathbb{N}})$.

The easiest way to produce a coupling is to let $\mathbb{P}_X$ and $\mathbb{P}_Y$ be any Markov chains over (S, P), and then let $\mathbb{P}_Z = \mathbb{P}_X \times \mathbb{P}_Y$ be the product measure on

$(S \times S)^{\mathbb{N}}$. This corresponds to running two copies of the Markov chain simultaneously and completely independently, without any interaction. However, this is far from being the only coupling available, and indeed there are other couplings that are far more useful for our applications. An important part of the power of the coupling method is that we can choose *any* distribution $\mathbb{P}_Z$ which has $\mathbb{P}_X$ and $\mathbb{P}_Y$ as its marginals – that is,

$$\mathbb{P}_X(A) = \mathbb{P}_Z(A \times S), \qquad \mathbb{P}_Y(A) = \mathbb{P}_Z(S \times A). \tag{6}$$

This allows us to introduce some dependence between X_n and Y_n, and in particular to carry out the following general scheme:

1. Define a coupling such that we eventually have $X_n = Y_n$ with probability 1 – that is, $\mathbb{P}_Z(X_n = Y_n) \to 1$ as $n \to \infty$.

2. Bound the total variation distance $d_V(p_x^m, \pi)$ that appears in (2) in terms of $\mathbb{P}_Z(X_m \neq Y_m)$, where X_m is a Markov chain starting in state x and Y_m starts in the stationary distribution π.

3. Use this bound to estimate the mixing time $\tau(\epsilon)$.

Before discussing how to bound total variation distance in terms of convergence times for couplings, we describe an example that illustrates how the two Markov chains X_n and Y_n can be chosen to have some dependence.

Let (S, P) be the Markov chain describing the top-down shuffle on a deck of n cards, which we discussed last time. That is, S is the set of all permutations of the n cards, and the transition probabilities are given by $p_{ij} = 1/n$ if j can be reached from i by removing a single card from the deck and placing it on top, and $p_{ij} = 0$ otherwise. One can define a coupling $Z_m = (X_m, Y_m)$ as follows: the Markov chains X_m and Y_m each progress from one step to the next by selecting a random card from the deck and placing it on top. Let Z_m progress from one step to the next by selecting a *single* card from the deck at random, and then moving that card to the top of the deck for both X_m and Y_m. Then it is clear that both X_m and Y_m evolve according to the transition probabilities p_{ij}, but the probability measure $\mathbb{P}_Z$ is not the direct product of $\mathbb{P}_X$ and $\mathbb{P}_Y$, because the chains do not evolve independently.

Note that it is the *card*, and not the position, which is the same between the two decks – in particular, after this process happens once, both decks X_m and Y_m have the same top card. After it happens twice, they have the same top *two* cards – unless at the second step we happen to pick the same card we did at the first. In general, we may write $a(m)$ for the number of cards at the top of the deck which we know to agree between X_m and Y_m. Then $a(0) = 0$ and $a(m)$ evolves according to the following rule: if a new card is picked at step m (that has not been picked before), then $a(m+1) = a(m)+1$, and otherwise $a(m+1) = a_m$.

In particular, we see that once every card in the deck has been chosen at least once, we have $a(m) = n$, so that $X_m = Y_m$. Later we will get an estimate on the probability that every card has been chosen by time m, which will let us estimate $\mathbb{P}_Z(X_m = Y_m)$. First we give an argument that this latter probability gives us a bound on the total variation distance.

3 Total variation distance and couplings

Ultimately we want to estimate $d_V(\mu_n, \nu_n)$, where $\mu_n = T^n(\mu)$ and $\nu_n = T^n(\nu)$ are the distributions at time n that result from the initial distributions μ and ν. The key is the following lemma.

Lemma 2 *Let X_n and Y_n be Markov chains over (S, P) with initial distributions μ and ν, respectively. Let $Z_n = (X_n, Y_n)$ be a coupling of X_n and Y_n. Then*

$$d_V(\mu_n, \nu_n) \leq \mathbb{P}_Z(X_n \neq Y_n). \tag{7}$$

Proof: Given any $A \subset S$, we have

$$\begin{aligned}
\mu_n(A) - \nu_n(A) &= \mathbb{P}_Z(X_n \in A) - \mathbb{P}_Z(Y_n \in A) \\
&= \mathbb{P}_Z(X_n \in A, Y_n \notin A) - \mathbb{P}_Z(Y_n \in A, X_n \notin A) \\
&\leq \mathbb{P}_Z(X_n \in A, Y_n \notin A) \\
&\leq \mathbb{P}_Z(X_n \neq Y_n).
\end{aligned}$$

□

In particular, if we take μ to be the initial distribution concentrated on a single state $x \in S$, and ν to be the stationary distribution π, then Lemma 2 gives us the estimate

$$d_V(p_x^m, \pi) \leq \mathbb{P}_Z(X_m \neq Y_m), \tag{8}$$

and so we have proved the following result.

Proposition 3 *Let Z_n be a coupling of Markov chains X_n, Y_n over (S, P), and let π be a stationary distribution for (S, P). Suppose that $T \in \mathbb{N}$ is such that for every $x \in S$, we have*

$$\mathbb{P}_Z(X_T \neq Y_T \mid X_0 = x, Y_0 \text{ distributed according to } \pi) \leq \epsilon.$$

Then $\tau(\epsilon) \leq T$.

4 Application to card shuffling

Now we can estimate the mixing time for the card shuffling example. Based on Proposition 3, we can estimate $\tau(\epsilon)$ by first estimating $\mathbb{P}_Z(X_T \neq Y_T)$, which as we saw above is the same as the probability that not every card has been selected by time T. The problem of determining how long it takes for this to happen is known as the coupon collector's problem.

If we run the Markov chain for T steps, then the probability that a specific card has not yet been selected to be moved to the top of the deck is

$$\left(1 - \frac{1}{n}\right)^T \approx e^{-T/n}.$$

Thus the probability that not every card has been selected is $\leq ne^{-T/n}$, and we get

$$\mathbb{P}_Z(X_T \neq Y_T) \leq ne^{-T/n}.$$

Note that this is independent of the starting distributions for X_0 and Y_0. We want to choose T such that this bound is $\leq \epsilon$, since then Proposition 3 will give $\tau(\epsilon) \leq T$. So we solve the inequality $ne^{-T/n} \leq \epsilon$ for T, and obtain

$$T \geq -n \log\left(\frac{\epsilon}{n}\right) = n \log n - n \log \epsilon,$$

which gives the rough bound

$$\tau(\epsilon) \leq n \log\left(\frac{n}{\epsilon}\right).$$

(Note that this bound depends on n being reasonably large.) For example, when $n = 52$ and $\epsilon = .05$, we get $\tau(\epsilon) \leq 361$, indicating that 361 shuffles (about 7 times the size of the deck) is enough to guarantee that for any event we specify, the probability of that event when drawing from our shuffled deck is within .05 of the probability when drawing from a truly random deck.