Lecture 2: October 7, 2003

Markov Chains, Coupling, Stationary Distribution

Eric Vigoda

2.1   Markov Chains

In this lecture, we will introduce Markov chains and show a potential algorithmic use of Markov chains for sampling from complex distributions.

For a finite state space Ω, we say a sequence of random variables (X_t) on Ω is a *Markov chain* if the sequence is Markovian in the following sense, for all t, all $x_0, \dots, x_t, y \in \Omega$, we require

$$\Pr(X_{t+1} = y | X_0 = x_0, X_1 = x_1, \dots, X_t = x_t) = \Pr(X_{t+1} = y | X_t = x_t).$$

We consider transitions which are independent of the time, known as time-homogeneous, and denote the transition matrix as

$$P(x, y) = \Pr(X_{t+1} = y | X_t = x).$$

The t-step distribution is defined in the natural way,

$$P^t(x, y) = \begin{cases} P(x, y) & t = 1 \\ \sum_{z \in \Omega} P(x, z) P^{t-1}(z, y) & t > 1 \end{cases}$$

We will study the class of ergodic Markov chains, which have a unique stationary (i.e., limiting) distribution and thus will be useful from an algorithmic perspective. We say a distribution π is a *stationary distribution* if it is invariant with respect to the transition matrix, i.e.,

$$\text{for all } y \in \Omega, \quad \pi(y) = \sum_{x \in \Omega} \pi(x) P(x, y). \tag{2.1}$$

A Markov chain is called *ergodic* if:

$$\text{there exists } t \text{ such that for all } x, y \in \Omega, \ P^t(x, y) > 0.$$

For finite Markov chains the following pair of conditions are equivalent to ergodicity:

- *Irreducible*: For all $x, y \in \Omega$, there exists $t = t(x, y)$ such that $P^t(x, y) > 0$;

- *Aperiodic*: For all $x \in \Omega$, $\gcd\{t : P^t(x, x) > 0\} = 1$.

Ergodic Markov chains are useful algorithmic tools in that, regardless of their initial state, they eventually reach a unique stationary distribution. We can aim to design (approximate) samplers by designing Markov chains with appropriate stationary distributions. The following theorem, originally proved by Doeblin [2], details the essential property of ergodic Markov chains.

Theorem 2.1 *For a finite ergodic Markov chain, there exists a unique stationary distribution π such that*

$$\text{for all } x, y \in \Omega, \quad \lim_{t \to \infty} P^t(x, y) = \pi(y).$$

Before proving the theorem, let us make a few remarks about its algorithmic consequences. In general, it is difficult to determine this unique stationary distribution. However, there is a large class of chains where it is trivial to verify the stationary distribution. A Markov chain is called *reversible* if there exists a distribution π such that

$$\text{for all } x, y \in \Omega, \quad \pi(x)P(x, y) = \pi(y)P(y, x).$$

It is easy to check that such a π is a stationary distribution. When P is symmetric, it follows that the stationary distribution is uniform over Ω.

The following example highlights the potential usefulness of reversible Markov chains. For a graph $G = (V, E)$, let Ω denote the set of matchings of G. We define a Markov chain on Ω whose transitions $X_t \to X_{t+1}$ are as follows. From $X_t \in \Omega$,

- Choose an edge e uniformly at random from E.

- Let

$$X' = \begin{cases} X_t \cup e & \text{if } e \notin X_t \\ X_t \setminus e & \text{if } e \in X_t \end{cases}$$

- If $X' \in \Omega$, then set $X_{t+1} = X'$ with probability 1/2; Otherwise set $X_{t+1} = X_t$.

Observe that the Markov chain is aperiodic (since $P(M, M) \geq 1/2$ for all $M \in \Omega$) and irreducible (via the empty set) with symmetric transition probabilities. Therefore, the unique stationary distribution is uniform over all matchings of G. If we can bound the asymptotic rate of convergence to the stationary distribution, then we have a simple algorithm to generate an approximately random matching. Simply start at an arbitrary matching (e.g., the empty set) and follow the transitions of the Markov chain until we are sufficiently close to the stationary distribution. For now we focus on proving the theorem. In later lectures we explore techniques for bounding the convergence rate.

2.2 Coupling Technique

We will prove the theorem using the *coupling* technique. We begin by describing coupling for general distributions before considering its application to Markov chains. For distributions μ, ν on a finite set Ω, a distribution ω on $\Omega \times \Omega$ is a *coupling* if:

$$\text{For all } x \in \Omega, \quad \sum_{y \in \Omega} \omega(x, y) = \mu(x), \text{ and} \qquad (2.2)$$

$$\text{For all } y \in \Omega, \quad \sum_{x \in \Omega} \omega(x, y) = \nu(y). \qquad (2.3)$$

In other words, ω is a joint distribution whose marginal distributions are the appropriate distributions.

Coupling provides a convenient method to bound the variation distance between a pair of distributions. For distributions μ and ν on Ω, their variation distance is defined as

$$d_{\text{TV}}(\mu, \nu) = \frac{1}{2} \sum_{z \in \Omega} |\mu(z) - \nu(z)| = \max_{S \subset \Omega} \mu(S) - \nu(S).$$

It turns out there always exists an optimal coupling which exactly captures the variation distance. The following lemma is known as the coupling lemma, and was first detailed by Aldous [1].

Lemma 2.2 *Consider a pair of distributions μ, ν on a finite Ω.*

(a) *For a coupling ω of μ and ν, let $(X, Y) \sim \omega$ (i.e., (X, Y) is a random variable chosen from the distribution ω). Then,*

$$d_{\text{TV}}(\mu, \nu) \leq \Pr(X \neq Y).$$

(b) *There always exists a coupling ω where, for $(X, Y) \sim \omega$,*

$$d_{TV}(\mu, \nu) = \Pr(X \neq Y).$$

Proof of Lemma, part (a):

Since ω is a valid coupling, for any $z \in \Omega$, we know that $\omega(z, z) \leq \min\{\mu(z), \nu(z)\}$. Summing over all z, this is exactly the probability that X and Y are equal, i.e.,

$$\Pr(X = Y) = \sum_{z \in \Omega} \omega(z, z) \leq \sum_{z \in \Omega} \min\{\mu(z), \nu(z)\}.$$

Therefore,

$$
\begin{aligned}
\Pr(X \neq Y) &\geq 1 - \sum_{z \in \Omega} \min\{\mu(z), \nu(z)\} \\
&= \sum_{z \in \Omega} \mu(z) - \min\{\mu(z), \nu(z)\} \\
&= \sum_{\substack{z \in \Omega: \\ \mu(z) \geq \nu(z)}} \mu(z) - \nu(z) \\
&= \max_{S \subset \Omega} \mu(S) - \nu(S) \\
&= d_{TV}(\mu, \nu).
\end{aligned}
$$

This completes the proof of part (a) of the lemma. ∎

Proof of Lemma, part (b):

For all $z \in \Omega$, let

$$
\omega(z, z) = \min\{\mu(z), \nu(z)\}.
$$

This ensures $d_{TV}(\mu, \nu) = \Pr(X \neq Y)$. Now we need to complete the construction of ω in a valid way. This requires defining the off-diagonal terms in a way that guarantees ω has the correct marginal distributions.

For $y, z \in \Omega, y \neq z$, let

$$
\omega(y, z) = \frac{(\mu(y) - \omega(y, y))(\nu(z) - \omega(z, z))}{1 - \sum_{x \in \Omega} \omega(x, x)}
$$

It is straightforward to verify that ω satisfies (2.2) and (2.3), thus it is a valid coupling. ∎

We will consider couplings for Markov chain. Consider a pair of Markov chains (X_t) and (Y_t) on Ω with transition matrices P_X and P_Y, respectively. Typically in our applications, the Markov chains are identical: $P_X = P_Y$. The Markov chain (X_t', Y_t') on $\Omega \times \Omega$ is a (Markovian) coupling if

$$
\begin{aligned}
\text{For all } a, b, c \in \Omega, \quad \Pr(X_{t+1}' = c | X_t' = a, Y_t' = b) &= P_X(a, c), \text{ and} \\
\text{For all } a, b, c \in \Omega, \quad \Pr(Y_{t+1}' = c | X_t' = a, Y_t' = b) &= P_Y(b, c).
\end{aligned}
$$

In other words, if we simply observe the first coordinate, it behaves like P_X and similarly the second coordinate acts according to P_Y. This is a more restrictive form of coupling then is necessary. In general, we might only have $\Pr(X_{t+1}' = c | X_t' = a) = P_X(a, c)$ and similarly for the second coordinate. Such a coupling is called a non-Markovian coupling, since the joint process is not itself a Markov chain on the product space. Most applications of coupling construct Markovian couplings, this simplifies the analysis.

For such a Markovian coupling (X'_t, Y'_t) of (X_t) and (Y_t) we then have

$$d_{TV}(P^t(X_0, \cdot), P^t(Y_0, \cdot)) \leq \Pr(X'_t \neq Y'_t | X'_0 = X_0, Y'_0 = Y_0).$$

Choosing Y_0 from the stationary distribution π, we have Y_t is distributed according to π for all π (since π is invariant), which implies

$$d_{TV}(P^t(X_0, \cdot), \pi) \leq \Pr(X'_t \neq Y'_t | X'_0 = X_0, Y'_0 \sim \pi).$$

This shows how we can use coupling to bound the distance from stationarity.

2.3 Stationary Distribution

We are now prepared to prove the theorem.

Proof of Theorem:

Create two copies of the Markov chain, denoted by (X_t) and (Y_t), where X_0 and Y_0 are arbitrary states of Ω. We will create a coupling for these chains in the following way. From (X_t, Y_t), choose X_{t+1} according to the transition matrix P. If $Y_t = X_t$, set $Y_{t+1} = X_{t+1}$, otherwise choose Y_{t+1} according to P, independent of the choice for X_t.

By ergodicity, we know that there exists t^* such that for all $x, y \in \Omega$, $P^{t^*}(x, y) \geq \epsilon > 0$. Therefore, for all $X_0, Y_0 \in \Omega$,

$$\Pr(X_{t^*} \neq Y_{t^*} \mid X_0, Y_0) \leq 1 - \epsilon.$$

Since this holds for all pairs of initial states, we can similarly look at the probability of coalescing during steps $t^* \to 2t^*$:

$$\Pr(X_{2t^*} \neq Y_{2t^*} | X_{t^*} \neq Y_{t^*}) \leq 1 - \epsilon.$$

Recall that, under our coupling, once $X_s = Y_s$ we have $X_{s'} = Y_{s'}$ for all $s' \geq s$. Therefore,

$$\Pr(X_{2t^*} \neq Y_{2t^*} \mid X_0, Y_0) = \Pr(X_{2t^*} \neq Y_{2t^*}, X_{t^*} \neq Y_{t^*} \mid X_0, Y_0)$$

Conditioning on not coalescing at time t^*, and then applying our earlier observation we have

$$\begin{aligned} \Pr(X_{2t^*} \neq Y_{2t^*} \mid X_0, Y_0) &= \Pr(X_{2t^*} \neq Y_{2t^*} | X_{t^*} \neq Y_{t^*}) \Pr(X_{t^*} \neq Y_{t^*} \mid X_0, Y_0) \\ &\leq (1 - \epsilon)^2 \end{aligned}$$

It is clear that for integer $k > 0$,

$$\Pr(X_{kt^*} \neq Y_{kt^*} \mid X_0, Y_0) \leq (1 - \epsilon)^k. \tag{2.4}$$

Therefore,
$$\Pr(X_{kt^*} \neq Y_{kt^*} \mid X_0, Y_0) \to 0 \ \text{ as } k \to \infty.$$

Since $X_t = Y_t$ implies $X_{t'} = Y_{t'}$ for all $t' \geq t$, we have

$$\Pr(X_t \neq Y_t \mid X_0, Y_0) \to 0 \ \text{ as } t \to \infty.$$

Note that the distributions of X_t and Y_t are $P^t(X_0, .)$ and $P^t(Y_0, .)$ respectively. The coupling of the Markov chain we defined, defines a coupling of the distributions $P^t(X_0, .)$ and $P^t(Y_0, .)$. Hence by Lemma 2.2

$$d_{TV}(P^t(X_0, \cdot), P^t(Y_0, \cdot)) \leq \Pr(X_t \neq Y_t) \to 0 \text{ as } t \to 0$$

This proves that from any starting points we approach the same distribution. Thus, for all $x, y \in \Omega$, we have

$$\lim_{t \to \infty} P^t(x, y) = \sigma(y),$$

i.e., there is a limiting distribution σ.

It remains to show that σ is a stationary distribution. Thus, we want to show that σ is an invariant distribution, i.e., it satisfies (2.1). For all $x, y \in \Omega$, we have

$$
\begin{aligned}
\sum_{x \in \Omega} \sigma(x) P(x, y) &= \sum_{x} \left(\lim_{t \to \infty} P^t(z, x) \right) P(x, y) \text{ for any } z \in \Omega \\
&= \lim_{t \to \infty} \left(\sum_{x} P^t(z, x) P(x, y) \right) \\
&= \lim_{t \to \infty} P^{t+1}(z, y) \\
&= \lim_{t \to \infty} P^t(z, y) \\
&= \sigma(y).
\end{aligned}
$$

Therefore, σ is an invariant distribution and it's a limiting distribution from any starting distribution, thus σ is the unique stationary distribution, which completes the proof of the theorem. ∎

2.4 Markov Chains for Algorithmic Purposes

For a Markov chain (e.g., the chain on matchings introduced earlier) to be useful for algorithmic purposes, we need that it converges quickly to its stationary distribution. The theorem we proved simply shows it converges in the limit over time, but it gives no indication as to the rate of convergence as a function of the size of the state space, i.e., $|\Omega|$. Therefore, we

define the *mixing time* $\tau_{\mathrm{mix}}(\epsilon)$ as the time until the chain is within variation distance ϵ from the worst initial state. Formally, let

$$\tau_{\mathrm{mix}}(\epsilon) = \max_{X_0 \in \Omega} \min\{t : d_{\mathrm{TV}}(P^t(X_0, \cdot), \pi) \leq \epsilon\}.$$

For the matchings chain, our hope is that $\tau_{\mathrm{mix}}(\epsilon) = poly(n, \log(1/\epsilon))$, where n is the number of vertices in the input graph. This then gives an efficient algorithm to approximately sample and approximately count matchings.

It suffices to reach variation distance below some constant, say $1/2e$. We can then boost to arbitrarily small variation distance, see the following exercise.

Exercise 2.3 *Prove*

$$\tau_{\mathrm{mix}}(\epsilon) \leq \tau_{\mathrm{mix}}(1/2e) \ln(1/\epsilon).$$

References

[1] D. J. Aldous. Random walks on finite groups and rapidly mixing Markov chains. In *Séminaire de Probabilites XVII*, pages 243–297. Springer Verlag, 1983. Lecture Notes in Mathematics 986.

[2] W. Doeblin. Exposé de la théorie des chaînes simples constantes de Markov á un nombre fini d'états. *Mathématique de l'Union Interbalkanique*, 2:77–105, 1938. Pages 78-80 reproduced in Lindvall [3].

[3] T. Lindvall. *Lectures on the coupling method*. John Wiley & Sons Inc., New York, 1992.