MARKOV CHAINS

ZSOLT TERDIK

Abstract. In this paper, I will buildup the basic framework of Markov Chains over finite state spaces using analytical techniques. In particular, we will see how the study of Markov Chains over finite state spaces reduces to the study of powers of matrices. Using this framework, I will prove that under mild restrictions, Markov Chains converge to a unique stationary distribution. Finally, I will discuss some interesting connections between Markov Chains and Linear Algebra.

Contents

1. Definitions and Preliminaries

For this paper, we are generally interested in a system that has finitely many *states*. We define the *state space* Ω as the finite set of possible states of our system. As a canonical example, the state space for a two sided coin is $\Omega_{coin} = \{Heads\ (H), Tails\ (T)\}$. Next we define an *event* $A \subseteq \Omega$ as a subset of the state space. Continuing with the coin example, possible events for three coins would be: $\{(H,H,H),(H,T,H),\}$ and $\{(H,T,T)\}$. Given a state space Ω, a finite Markov Chain moves, or *transitions*, between elements in the space in discrete times steps according to a fixed probability distribution $P(x,\cdot)$ for all $x \in \Omega$. In this view, a Markov Chain is a sequence of random of random variable $X_1, X_2, \dots$ taking on values in Ω. The defining property of Markov Chains is that the transition probability is *fixed* for each state $x \in \Omega$, and hence not dependent on previous states. Hence, we have for all $x \in \Omega$, $t \in \mathbb{N}$:

$$\mathbf{P}\{X_{t+1} = y | X_0 = x_0, \dots, X_t = x_t\} = \mathbf{P}\{X_{t+1} = y | X_t = x_t\}. \tag{1.1}$$

From Equation 1.1, we see that any finite Markov Chain is completely specified by its transition matrix $P \in \mathbf{Mat}_{|\Omega|\times|\Omega|}$ where $P_{x,y}$ is the transition probability between states $x \to y$. Given that we require the matrix to be *stochastic*, that is for any state $x \in \Omega$, we want the sum of the probability distribution over the entire state space to be 1. That is, for all $x \in \Omega$:

$$\sum_{y \in \Omega} P(x, y) = 1.$$

In terms of matrices we require that all entries be non-negative, and the row sums to be 1. In general however, the matrix need not be symmetric. Examining this matrix we see that the k-th row is the distribution associated with transitioning from the k-th state to any state, and similarly the j-th column is the probability of transitioning from any state to the j-th state.

We are now ready to build up a computational framework. Suppose we have a Markov Chain with transition matrix P. We define a probability distribution on Ω as a *row vector* where the i-th entry is the probability associated with the i-th state. Analogously, define a *function* as a *column vector.* In general, we define the *length* of distribution, $|\pi|$ as the number of coordinates; similarly, we define the size of the state space, $|\Omega|$ as the number of possible states. Consider for example an even distribution π_{even} across on the state space $\Omega = \{x_o, x_1, x_2, \ldots, x_{n-1}\}$:

$$\pi_{even} = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right), \quad |\pi_{even}| = n,$$

and a distribution weighted completely at x_1:

$$\pi_{x_1} = (0, 1, 0, \ldots, 0).$$

More generally, a distribution π is a row vector with entries $\pi_i = p(x_i)$ It is not difficult to see that a distribution π multiplied by a function f will yield the expected value of the function with respect to π. In terms of matrices we have,

$$\pi \cdot f = (\pi^1, \pi^2, \cdots, \pi^{|\Omega|}) \cdot (f_1, f_2, \cdots, f_{|\Omega|}) = \sum_{i=1}^{|\Omega|} \pi^i f_i = \mathbf{E}_\pi(f). \tag{1.2}$$

Diving deeper into the matrix machinery, we see that multiplying an initial distribution π_0 on the right by a transition matrix P will yield "tomorrow's" distribution. Extending this result to finite powers, we have:

$$\pi_0 \cdot P^t = (\pi_0 P) \cdot P^{t-1} = \pi_1 \cdot P^{t-1} = \cdots = \pi_t \cdot P^0 = \pi_t. \tag{1.3}$$

That is, the transition matrix specifies the time evolution of a probability distribution. We insert the transition matrix between a row and column vector to obtain:

$$(\pi_0 P^t) f = \pi_t f = \mathbf{E}_{\pi_t}(f). \tag{1.4}$$

Evaluating what happens to a distribution and function under a finite number of transitions is then reduced to evaluating finite powers of the associated transition matrix. However, the natural question becomes what happens to π_0 (or f) as $t \to \infty$?

2. Convergence to Stationary Distributions

Before we begin, we will need to classify a few basic distributions and Markov Chains. We say that a distribution π is *stationary* with respect to transition matrix P if:

$$\pi \cdot P = \pi. \tag{2.1}$$

We say a Markov Chain is *irreducible* if for any points $x, y \in \Omega$ there exists $n \in \mathbb{N}$ such that:

$$P^n(x, y) > 0. \tag{2.2}$$

The notion of irreducibility has a particularly clean interpretation if we consider a weighted directed graph $G = \{\Omega, E\}$ with the vertex set as the state space, and with edges between states of positive transition probability, and edge weights of the corresponding transition probability. The notion of irreducibility corresponds to connectedness of the associated graph. That is, there exists a finite path connecting any two vertices, if and only if the Markov Chain is irreducible. Furthermore we state a basic result as Proposition , the proof of which may be found in [1].

Proposition 2.3. *If P is finite, irreducible and aperiodic, then there exists a integer r such that $P^r(x, y) > 0$ for all $x, y \in \Omega$, that is for any Markov Chain under the usual assumption, some power of the transition matrix will have strictly positive entries.*

Finally, we define total variation distance, which gives us a metric on the space of distributions.

Definition 2.4 (Metric on space of Distirbutions)**.** Given two probability distribution π and μ we define the *total variation distance* as:

$$\|\pi - \mu\|_{TV} := \max_{A \subset \Omega} |\pi(A) - \mu(A)|. \tag{2.5}$$

We may prove an alternate formulation of 2.4 as:

$$||\pi - \mu||_{TV} = \frac{1}{2} \sum_{x \in \Omega} |\pi(x) - \mu(x)|. \tag{2.6}$$

To prove equivalence, consider the complementary events:

$$\begin{aligned} B^+ &:= \{x \in \Omega : \pi(x) - \mu(x) \geq 0\} \\ B^- &:= \{x \in \Omega : \pi(x) - \mu(x) < 0\}. \end{aligned}$$

Fix $A \subset \Omega$. Consider:

$$(\pi - \mu)(A) = \pi(A) - \mu(A) \leq \pi(A \cap B^+) - \mu(A \cap B^+) \leq \pi(B^+) - \mu(B^+), \tag{2.7}$$

where the first inequality follows since $\pi - \mu$ is positive on B^+. Applying the analogous logic to $(\mu - \pi)(A)$, we have:

$$(\mu - \pi)(A) \leq \mu(B^-) - \pi(B^-). \tag{2.8}$$

We also see immediately that if $A = B^+$ or $A = B^-$ then the corresponding upper bounds are sharp, and hence the $\max_{A \subset \Omega}$ is obtained precisely for $A = B^+$ and $A = B^-$. Now, since $B^+ \sqcup B^- = \Omega$, we may add the inequalities and divide the result by two:

$$\max_{A \subset \Omega} |\pi(A) - \mu(A)| = \frac{1}{2}\big[\pi(B^+) - \mu(B^+) + \mu(B^-) - \pi(B^-)\big] = \frac{1}{2} \sum_{x \in \Omega} |\pi(x) - \mu(x)|. \tag{2.9}$$

Hence, we have that total variation distance corresponds to the maximum difference in *area* between the two distribution.

Theorem 2.10. *Let P be the transition matrix associated with an irreducible Markov Chain over a finite state space Ω. Then:*

1. *there exists a stationary probability distribution π,*
2. *the stationary distribution is unique, and*
3. *(Convergence) there exists constants $\alpha \in (0,1)$ and $C > 0$ such that:*

$$\max_{x \in \Omega} \|P^t(x, \cdot) - \pi\|_{TV} \leq C\alpha^t.$$

Proof. (1) First, fix $x \in \Omega$ and let μ be an arbitrary distribution. Now consider the distribution formed by taking $n > 0$:

$$\mu_n(x) := \frac{1}{n}(\mu(x) + (\mu P)(x) + \cdots + (\mu P^{n-1})(x)) = \frac{1}{n}\sum_{t=0}^{n-1}(\mu P^t)(x).$$

This probability distribution is well defined. In particular it is stochastic. To show this, consider a simple average of two distributions:

$$\frac{1}{2}\sum_{x\in\Omega}(\pi + \mu)(x) = \frac{1}{2}\Big[\sum_{x\in\Omega}\pi(x) + \sum_{x\in\Omega}\mu(x)\Big] = \frac{1}{2}\cdot[1+1] = 1.$$

This will obviously generalize to an average of n distributions. We now want to find a bound on the change after an application of P. Hence, consider

(2.11)

$$|(\mu_n P)(x) - \mu_n(x)| = \frac{1}{n}\left|\sum_{t=1}^{n}(\mu P^t)(x) - \sum_{t=0}^{n-1}(\mu P^t)(x)\right| = \frac{1}{n}|(\mu P^n)(x) - \mu(x)| \leq \frac{1}{n}.$$

Where the inequality follows from the fact that we have a probability measure and hence the difference between any two probabilities is at most 1. Now that we have our candidate, bounded sequence, we may apply Bolzano-Weistrauss, which states that any bounded sequence in $\mathbb{R}^n$ has a convergent sub-sequence. Here we use the fact that $|\Omega| < \infty$ and that all terms in the row vector corresponding the distribution are necessarily finite. We take our candidate stationary distribution π to be the pointwise limit of the convergent sub-sequence μ_{n_k}. That is:

$$\pi(x) := \lim_{k\to\infty}\mu_{n_k}(x).$$

We may now check explicitly that our candidate distribution is stochastic and stationary. First, to show that the distribution is stochastic note that we can interchange limits and finite sums. Consider:

$$\sum_{x\in\Omega}\pi(x) = \sum_{x\in\Omega}\lim_{k\to\infty}\mu_{n_k} = \lim_{k\to\infty}\sum_{x\in\Omega}\mu_{n_k} = \lim_{k\to\infty} 1 = 1.$$

Now to show π is stationary, fix $x \in \Omega$. Consider:

$$\pi(x)\cdot P - \pi(x) = \lim_{n_k\to\infty}\mu_{n_k}(x)\cdot P - \lim_{n_k\to\infty}\mu_{n_k}(x) = 0,$$

by 2.5 and the condition that $n_k \to \infty$

(2) We are now in a position to prove directly that the stationary distribution is unique. Suppose, by contradiction there exists two stationary distribution π_1 and π_2. Choose any $x \in \Omega$. Since the matrix is irreducible, we have

strictly positive transition probabilities. That is, for any $y \in \Omega$, $P^t(x, y) > 0$ for some $t > 0$. First, choose $y \in \Omega$ such that $P(x, y) > 0$. Hence consider:

$$\frac{\pi_1(x)}{\pi_2(x)} = \frac{\pi_1(x)}{\pi_2(x)} \cdot \frac{P(x,y)}{P(x,y)} = \frac{\pi_1(y)}{\pi_2(y)}, \tag{2.12}$$

where the second equality follows from the fact that both π_1 and π_2 are stationary. It is important to note at this point that a stationary distribution must have strictly positive entries which follows immediately from the irreducibility assumption. Similarly by the irreducibility assumption, we have that this result holds for all points $x, y \in \Omega$, and equality of the stationary distribution is established transitively by equating neighboring points. This logic is succinctly stated as:

$$\frac{\pi_1(x)}{\pi_2(x)} = \frac{\pi_1(x)}{\pi_2(x)} \cdot \frac{P^t(x,z)}{P^t(x,z)} = \frac{\pi_1(z)}{\pi_2(z)}, \tag{2.13}$$

for some $t > 0$ and all $z \in \Omega$. Hence, we have for any state x, that the ratio between the distributions is a constant, c. That is:

$$\pi_1(x) = c \cdot \pi_2(x). \tag{2.14}$$

Since any distribution must be stochastic, summing over all states $x \in \Omega$ we have:

$$1 = \sum_{x \in \Omega} \pi_1(x) = \sum_{x \in \Omega} c \cdot \pi_2(x) = c \cdot \sum_{x \in \Omega} \pi_2(x) = c. \tag{2.15}$$

Since $c = 1$, we have that that distributions are equal on all states.

(3) Having shown existence (1) and uniqueness (2), we want to show that any initial distribution will in fact converge to the unique stationary distribution. This proof will in general proceed by decomposing the transition matrix P into two matrices Π and Q where Π will move the distribution directly to the stationary distribution, and Q is just some matrix which formally encodes the probability of *not* moving directly to stationary distribution. It is not difficult to see, directly from the definition, that once a stationary distribution is achieved, the Markov chain *will remain stationary.* Since Π was constructed to move directly to stationary, we will want to study the probability of picking Q by randomly choosing between Π and Q. In particular, in order to show convergence we want the probability of never picking Π to decay sufficiently fast as $t \to \infty$.

To proceed with the central theorem, let π be the stationary distribution, and let $\Pi \in \mathbf{Mat}_{|\Omega| \times |\Omega|}$ such that every row is the stationary distribution π. We have from Proposition (2.8), there exists a sufficiently small $\delta > 0$ such that:

$$P^r(x, y) \geq \delta \cdot \pi(y),$$

for all $x, y \in \Omega$. We now define the matrix Q to satisfy the following:

$$P^r := (1 - \theta)\Pi + \theta Q, \tag{2.16}$$

where $\theta := 1 - \delta$. We may check that Q is in fact a stochastic matrix, that is every row sums to one. Note that Π is clearly a stochastic matrix, since every row is a distribution, and P^r is also stochastic since it is positive power of the transition matrix P. Fix $x \in \Omega$. Summing over the columns, we obtain the row-sum of Q:

$$\sum_{y\in\Omega} Q(x,y) = \sum_{y\in\Omega} \frac{P^r(x,y) - (1-\theta)\Pi(x,y)}{\theta} = \frac{1-(1-\theta)}{\theta} = 1.$$

Now we prove by induction with respect to k that:

$$P^{rk} = (1-\theta^k)\Pi + (\theta Q)^k. \tag{2.17}$$

We start with that base case $k = 1$ which is true by from the definition of Q in (2.9). Now assume (2.10) hold for $k = n$. For our induction step:

$$\begin{aligned}
P^{r(n+1)} &= P^{rn} \cdot P^r = \Big((1-\theta^n)\Pi + \theta^n Q^n\Big)P^r \\
&= (1-\theta^n)\Pi \cdot P^r + \theta^n Q^n P^r \\
&= (1-\theta^n)\Pi \cdot P^r + \theta^n Q^n \cdot \big((1-\theta)\Pi + \theta Q\big) \\
&= (1-\theta^n)\Pi \cdot P^r + \theta^n (1-\theta) Q^n \Pi + \theta^{n+1} Q^{n+1} \\
&= (1-\theta^n)\Pi + \theta^n (1-\theta)\Pi + \theta^{n+1} Q^{n+1} \\
&= (1-\theta^n + \theta^n - \theta^{n+1})\Pi + \theta^{n+1} Q^{n+1} \\
&= (1-\theta^{n+1})\Pi + (\theta Q)^{n+1}.
\end{aligned}$$

Hence, we have shown 2.10 for all k. Now, multiplying 2.10 through by P^j so that our result will include for all powers of P, not just multiples of k, we obtain:

$$P^{rk+j} - \Pi = \theta^k (Q^k P^j - \Pi).$$

Now, in order to compute $\|\cdot\|_{TV}$ we fix $x \in \Omega$ and sum the left and multiply by $1/2$ to obtain:

$$\frac{1}{2}\sum_{y\in\Omega} |P^{rk+j}(x,y) - \Pi(x,y)| = \|P^{rk+j}(x,\cdot) - \pi\|_{TV}. \tag{2.18}$$

Computing the total variation distance for the right hand side, we note that:

$$\|Q^k P^j(x,\cdot) - \pi\| \leq 1,$$

and hence we may bound (2.11) from above to obtain the desired result:

$$\|P^{rk+j}(x,\cdot) - \pi\|_{TV} \leq \theta^k. \tag{2.19}$$

To summarize, this shows that by decomposing the transition matrix (P) into a matrix of stationary distributions (Π) and some other residue matrix Q, we may force the total variation distance between distributions at time t and the stationary distribution to decrease exponentially with respect to t.

□

3. Markov Chains and Linear Algebra

Consider an n-gon W inscribed in the unit circle on $\mathbb{C}$. Now, let the state space Ω be the vertices:

$$\Omega := \{1, w, w^2, \ldots, w^{n-1}\}.$$

Consider a Markov Chain that transitions between adjacent vertices with equal probability. The transition matrix will be:

$$P_{x,y} = \begin{cases} \frac{1}{2}, & \text{if } x \text{ is adjacent to } y \\ 0, & \text{else.} \end{cases}$$

This corresponds to a simple random walk on the n-gon. We see, from the definition of eigenfunction, that the following must be satisfied:

$$Pf(w^k) = \lambda \cdot f(w^k). \tag{3.1}$$

Since the probability of transitioning to each of the neighbors w^{k-1} and w^{k+1} is 1/2, we have:

$$\lambda \cdot f(w^k) = \frac{f(w^{k-1}) + f(w^{k+1})}{2}. \tag{3.2}$$

To find the eigenvalues, consider:

(3.3)

$$P(w^{jk}) = (Pw^j)(w^k) = \frac{w^j(w^{k-1}) + w^j(w^{k+1})}{2} = \frac{w^{jk-j} + w^{jk+j}}{2} = w^{jk}\left(\frac{w^{-j} + w^j}{2}\right).$$

Hence, $\frac{w^{-j}+w^j}{2} = \cos(2\pi j/n)$ is an eigenvalue of w^{jk}. However, this has a simple geometric interpretation as the length of the projection of tomorrow's vector w^{j+l} onto the initial vector w^j.

In general, linear algebra proves to versatile tool in study of Markov Chains precisely because a Markov Chain may be concretely represented as a transition matrix. Finding the unique stationary distribution then amounts to finding the eigenvector π with eigenvalue $\lambda = 1$. The second and higher eigenvalues may be used to classify Markov Chains through *spectral gap* analysis. Furthermore, one may obtain robust lower and upper bounds by analyzing the spectral gap. For more information see [1], in particular Chapters 12 and 13.

Acknowledgments. It is a pleasure to thank my mentors, Brent and Tomasz for their patience, clear insights, and impossibly expedient late night email responses.

References

[1] Levin, Peres, and Wilmer. Markov Chains and Mixing Times. American Mathematical Society. 2009.