CONVERGENCE THEOREM FOR FINITE MARKOV CHAINS

ARI FREEDMAN

ABSTRACT. In this expository paper, I will give an overview of the necessary conditions for convergence in Markov chains on finite state spaces. In doing so, I will prove the existence and uniqueness of a stationary distribution for irreducible Markov chains, and finally the Convergence Theorem when aperiodicity is also satisfied.

CONTENTS

1. INTRODUCTION AND BASIC DEFINITIONS

A Markov chain is a stochastic process, i.e., randomly determined, that moves among a set of states over discrete time steps. Given that the chain is at a certain state at any given time, there is a fixed probability distribution for which state the chain will go to next (including repeating the state). If there are n states, then the $n \times n$ transition matrix P describes the Markov chain, where the rows and columns are indexed by the states, and $P(x, y)$, the number in the x-th row and y-th column gives the probability of going to state y at time $t + 1$, given that it is at state x at time t. We can formalize this as follows.

Definition 1.1. A **finite Markov chain** with finite **state space $\boldsymbol{\Omega}$** and $|\Omega| \times |\Omega|$ transition matrix P is a sequence of random variables $X_0, X_1, \ldots$ where

$$\mathbf{P}\{X_{t+1} = y \mid X_t = x\} = P(x, y),$$

or the probability of $X_{t+1} = y$ given $X_t = x$ is $P(x, y)$. Then $P(x, \cdot)$, the x-th row of P, gives the distribution of X_{t+1} given $X_t = x$. Here $\mathbf{P}$ is the notation for probability of an event, and $\mathbf{P}_x$ notates the probability of an event given $X_0 = x \in \Omega$. Thus, $\mathbf{P}_x\{X_t = y\} = \mathbf{P}\{X_t = y \mid X_0 = x\} = P^t(x, y)$, as multiplying a distribution by the transition matrix P advances the distribution one step along the Markov chain, so multiplying by P^t advances it by t steps from $X_0 = x$.

Furthermore, when the distribution of X_{t+1} is conditioned on X_t, the previous values in the chain, $X_0, X_1, \ldots, X_{t-1}$, do not affect the value of X_{t+1}. This is called

Date: August 28, 2017.

the **Markov property**, and can be formalized by saying that if $H = \bigcap_{i=0}^{t-1}\{X_i = x_i\}$ is any event such that $\mathbf{P}(H \cap \{X_t = x\}) > 0$, then

$$P\{X_{t+1} = y \mid H \cap \{X_t = x\}\} = P\{X_{t+1} = y \mid X_t = x\},$$

and then we simply define $P(x, y) = P\{X_{t+1} = y \mid X_t = x\}$.

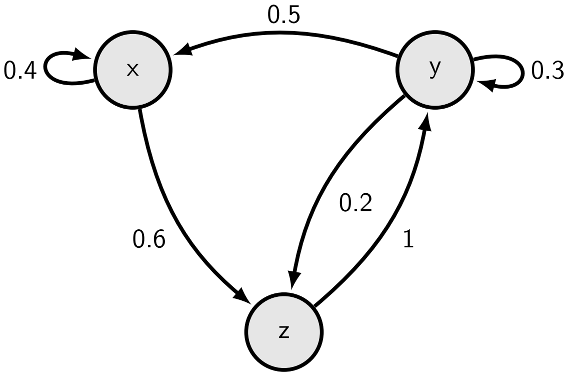

Figure 1. An example Markov chain with three states x, y, and z.

We can illustrate a Markov chain with a state diagram, in which an arrow from one state to another indicates the probability of going to the second state given we were just in the first. For example, in this diagram, given that the Markov chain is currently in x, we have probability .4 of staying in x, probability .6 of going to z, and probability 0 of going to y in the next time step (Fig. 1). This Markov chain would be represented by the transition matrix

$$P = \begin{array}{c c} & \begin{array}{ccc} x & y & z \end{array} \\ \begin{array}{c} x \\ y \\ z \end{array} & \begin{pmatrix} .4 & 0 & .6 \\ .5 & .3 & .2 \\ 0 & 1 & 0 \end{pmatrix} \end{array}.$$

This definition mentions distributions, so it may help to formalize what these are.

Definition 1.2. A **probability distribution**, or just a **distribution**, is a vector of non-negative probabilities that sums up to 1 (this is known as the law of total probability).

For any state $x \in \Omega$, it makes sense that $P(x, \cdot)$, the x-th row of P, should be a distribution, since the probability of going from state x to any state is at least 0, and the sum of the probabilities of going from state x to state y, over all states $y \in \Omega$, should be 1, as these are disjoint events that cover all the possibilities. Distributions are generally expressed as row vectors, which can then be right-multiplied by matrices.

The transition matrices associated with Markov chains all fall under the larger category of what we call stochastic matrices.

Definition 1.3. A **stochastic matrix** is an $n \times n$ matrix with all non-negative values and each row summing to 1. In particular, a matrix is stochastic if and only if it consists of n distribution row vectors in $\mathbb{R}^n$.

It is fairly easy to see that if P and Q are both stochastic matrices, then PQ is also a stochastic matrix, and if μ is a distribution, then μP is also a distribution.

Definition 1.4. A distribution π is called a **stationary distribution** of a Markov chain P if $\pi P = \pi$.

Thus, a stationary distribution is one for which advancing it along the Markov chain does not change the distribution: if the distribution of X_t is a stationary distribution π, then the distribution of X_{t+1} will also be π. This brings up the question of when a Markov chain will have a stationary distribution, and if so, is this distribution unique? Will any distribution converge to this stationary distribution over time? It turns out that with only mild constraints, all of these are satisfied.

Definition 1.5. A Markov chain is **irreducible** if for all states $x, y \in \Omega$, there exists a $t \geq 0$ such that $P^t(x, y) > 0$.

Intuitively, this means that it is possible to get from x to y for any $x, y \in \Omega$ in some finite amount of time steps, or, equivalently, there exists a sequence of states $x = x_0, x_1, \ldots, x_{t-1}, x_t = y$ (which we call a **path**) the chain can take from x to y such that $P(x_i, x_{i+1}) > 0$ for all $0 \leq i < t$.

2. Uniqueness of Stationary Distributions

In order to come up with a nice expression to show the existence of a stationary distribution, we must first prove that if a distribution exists, it is unique. To do so, we need to define harmonic functions and prove a lemma about them.

Definition 2.1. A function $h : \Omega \to \mathbb{R}$ is a **harmonic at** $x \in \Omega$ if

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

If h is harmonic at all states in $\Omega = \{x_1, x_2, \ldots, x_n\}$, we say h is harmonic on Ω, and then $Ph = h$, where h is the column vector $h = \begin{pmatrix} h(x_1) \\ h(x_2) \\ \vdots \\ h(x_n) \end{pmatrix}$.

Lemma 2.2. *If P is irreducible and h is harmonic on Ω, then h is a constant function.*

Proof. Since Ω is finite, h attains a maximum at some state $x_0 \in \Omega$, such that $h(x_0) \geq h(y)\ \forall y \in \Omega$. Let $z \in \Omega$ be any state such that $P(x_0, z) > 0$, and assume that $h(z) < h(x_0)$. Since h is harmonic at x_0,

$$\begin{aligned}
h(x_0) &= \sum_{y \in \Omega} P(x_0, y) h(y) \\
&= P(x_0, z) h(z) + \sum_{y \in \Omega, y \neq z} P(x_0, y) h(y) \\
&\leq P(x_0, z) h(z) + \sum_{y \in \Omega, y \neq z} P(x_0, y) h(x_0) \\
&< P(x_0, z) h(x_0) + \sum_{y \in \Omega, y \neq z} P(x_0, y) h(x_0) \\
&= \left(\sum_{y \in \Omega} P(x_0, y) \right) h(x_0) \\
&= h(x_0),
\end{aligned}$$

where the last inequality follows from $P(x_0, z) > 0$ and $h(z) < h(x_0)$. However, this gives us $h(x_0) < h(x_0)$, a contradiction, which means $h(x_0) = h(z)$.

Now for any $y \in \Omega$, P being irreducible implies there is a path from x_0 to y, let it be $x_0, x_1, \ldots, x_n = y$ such that $P(x_i, x_{i+1}) > 0$. Thus, $h(x_0) = h(x_1)$, and so x_1 also maximizes h, which means $h(x_1) = h(x_2)$. We carry on this logic to get

$$h(x_0) = h(x_1) = \cdots = h(x_n) = h(y).$$

So $h(y) = h(x_0)$ $\forall y \in \Omega$, and thus h is a constant function. □

Now we are ready to show that if a stationary distribution exists (which we show in the next section), it must be unique.

Corollary 2.3. *If P is irreducible and has a stationary distribution π, then π is the only such stationary distribution.*

Proof. By Lemma 2.2, the only functions h that are harmonic are those of the form $h(x) = c$ $\forall x \in \Omega$ for some constant c. Putting this into vector form, this means the only solutions to the equation $Ph = h$, or equivalently $(P - I)h = 0$ are

$$h = c \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Thus, $\dim(\ker(P - I)) = 1$, so by the rank-nullity theorem, $\operatorname{rank}(P - I) = |\Omega| - 1$. And $\operatorname{rank}(P - I) = \operatorname{rank}((P - I)^T) = \operatorname{rank}(P^T - I) = |\Omega| - 1$, so again by the rank-nullity theorem, $\dim(\ker(P^T - I)) = 1$, so over all row vectors $v \in \mathbb{R}^{|\Omega|}$, the equation $(P^T - I)v^T = 0$ has only a one-dimensional space of solutions. But this equation is equivalent to $vP = v$, so, given that $\pi P = \pi$ is a solution, all solutions must be of the form $v = \lambda \pi$, for some scalar λ. However, to be a distribution whose elements sum to 1, we must have $\lambda = 1$, and thus the only stationary distribution is $v = \pi$. □

3. Existence of a Stationary Distributions

We will now show that all irreducible Markov chains have a stationary distribution by explicitly constructing one, and then by Corollary 2.3, we will know that this stationary distribution is unique.

Definition 3.1. For a Markov chain $X_0, X_1, \ldots$, the **hitting time** for a state $x \in \Omega$ is the instance of the chain "hitting" x, notated

$$\tau_x = \min\{t \geq 0 : X_t = x\}.$$

When we want the hitting time to be strictly positive, we notate it

$$\tau_x^+ = \min\{t > 0 : X_t = x\},$$

which is called the **first return time** when $X_0 = x$.

We will also be using the notation $\mathbf{E}$ to denote the expected value of a variable, and again, $\mathbf{E}_x$ means the expected value given $X_0 = x$.

Lemma 3.2. *For any $x, y \in \Omega$ of an irreducible Markov chain, $\mathbf{E}_x(\tau_y^+)$ is finite.*

Proof. Since P is irreducible, we know for any two states $z, w \in \Omega$ that there exists an $s > 0$ such that $P^s(z, w) > 0$ (if $z = w$, we consider a path from z to a different state and back to itself to ensure $s > 0$). We let r be the maximum of all such s over $z, w \in \Omega$, and let

$$\epsilon = \min\{P^s(z, w) > 0 \text{ such that } 0 < s \leq r : z, w \in \Omega\}.$$

Then for all $z, w \in \Omega$, there exists $0 < s \leq r$ such that $P^s(z, w) \geq \epsilon > 0$.

This implies that given any X_t, the probability of the chain going to a state y between times t and $t + r$ is at least ϵ, or conversely,

$$\mathbf{P}\{X_s \neq y : \forall t < s \leq t + r\} \geq 1 - \epsilon.$$

In general, saying $\tau_y^+ > n$ implies $X_t \neq y \ \forall 0 < t \leq n$. So for $k > 0$,

$$\begin{aligned}
\mathbf{P}_x\{\tau_y^+ > kr\} &= \mathbf{P}_x\{X_t \neq y \ \forall 0 < t \leq kr\} \\
&= \mathbf{P}_x\{X_t \neq y \ \forall 0 < t \leq (k-1)r\}\mathbf{P}_x\{X_t \neq y \ \forall (k-1)r < t \leq kr\} \\
&\leq \mathbf{P}_x\{X_t \neq y \ \forall 0 < t \leq (k-1)r\}(1 - \epsilon) \\
&\leq \mathbf{P}_x\{X_t \neq y \ \forall 0 < t \leq (k-2)r\}(1 - \epsilon)^2 \\
&\ \ \vdots \\
&\leq \mathbf{P}_x\{X_t \neq y \ \forall 0 < t \leq 0\}(1 - \epsilon)^k \\
&= (1 - \epsilon)^k,
\end{aligned}$$

where $\mathbf{P}_x\{X_t \neq y \ \forall 0 < t \leq 0\} = 1$ by vacuous truth.

Now for any random variable Y valued on the non-negative integers, we have

$$\begin{aligned}
E(Y) &= \sum_{t=0}^{\infty} tP(Y = t) \\
&= 1 \cdot P(Y = 1) + 2 \cdot P(Y = 2) + 3 \cdot P(Y = 3) + \cdots \\
&= \Big(P(Y = 1) + P(Y = 2) + P(Y = 3) + \cdots\Big) + \Big(P(Y = 2) + P(Y = 3) + \cdots\Big) + \cdots \\
&= \sum_{t=0}^{\infty} P(Y > t).
\end{aligned}$$

And $\mathbf{P}\{\tau_y^+ > t\}$ is a decreasing function with respect to t, since

$$\mathbf{P}\{\tau_y^+ > t+1\} \leq \mathbf{P}\{\tau_y^+ > t+1\} + \mathbf{P}\{\tau_y^+ = t+1\} = \mathbf{P}\{\tau_y^+ > t\},$$

which we use to get

$$\begin{aligned}\mathbf{E}_x(\tau_y^+) &= \sum_{t=0}^{\infty} \mathbf{P}_x\{\tau_y^+ > t\} \\ &\leq \sum_{k=0}^{\infty} r\mathbf{P}_x\{\tau_y^+ > kr\} \\ &\leq r\sum_{k=0}^{\infty}(1-\epsilon)^k.\end{aligned}$$

By definition, $0 < \epsilon \leq 1$, so $0 \leq 1-\epsilon < 1$, which means this sum converges, and thus $\mathbf{E}_x(\tau_y^+)$ is finite. □

Theorem 3.3. *If P is irreducible, then it has a unique stationary distribution π with $\pi(x) > 0$ $\forall x \in \Omega$, given by*

$$\pi(x) = \frac{1}{\mathbf{E}_x(\tau_x^+)}.$$

Proof. Fix any state $z \in \Omega$. Then we define

$$\begin{aligned}\widetilde{\pi}(y) &= \mathbf{E}_z(\text{number of visits to } y \text{ before returning to } z) \\ &= \sum_{t=0}^{\infty} \mathbf{P}_z\{X_t = y, \tau_z^+ > t\},\end{aligned}$$

since the expected number of visits to y before returning to z is the sum of all the probabilities of the chain hitting y at a time step less than the return time.

For any given chain with $X_0 = z$, the number of visits to y before return to z is $\leq \tau_z^+$, since the total number of states the chain visits before returning to z is τ_z^+. Thus $\widetilde{\pi}(y) \leq \mathbf{E}_z(\tau_z^+)$, which by Lemma 3.2 is finite, and thus all $\widetilde{\pi}(y)$ are finite.

And since P is irreducible, it is at least possible to visit y once (a path from z to y followed by a path from y to z), which means the expected number of visits to y before returning to z is positive, so $\widetilde{\pi}(y) > 0$.

Now we show $\widetilde{\pi}$ is stationary, or that for all y, $(\widetilde{\pi}P)(y) = \widetilde{\pi}(y)$. First, see that

$$(\widetilde{\pi}P)(y) = \sum_{x\in\Omega} \widetilde{\pi}(x)P(x,y) = \sum_{x\in\Omega}\sum_{t=0}^{\infty} \mathbf{P}_z\{X_t = x, \tau_z^+ \geq t+1\}P(x,y),$$

where we just plugged in our earlier expression for $\widetilde{\pi}(x)$ and replaced $\tau_z^+ > t$ with the equivalent expression $\tau_z^+ \geq t+1$.

Since the event $\{\tau_z^+ \geq t+1\}$ is only determined by $X_0, X_1, \ldots, X_t$, it is independent of the event $X_{t+1} = y$, when conditioned on $X_t = x$, which means

$$\begin{aligned}\mathbf{P}_z\{X_t = x, X_{t+1} = y, \tau_z^+ \geq t+1\} &= \mathbf{P}_z\{X_t = x, \tau_z^+ \geq t+1\}\mathbf{P}_z\{X_{t+1} = y \mid X_t = x\} \\ &= \mathbf{P}_z\{X_t = x, \tau_z^+ \geq t+1\}P(x,y).\end{aligned}$$

We can then plug this in to our earlier expression for $(\widetilde{\pi}P)(y)$ and switch around the order of summation, since the inner sum converges for all $x \in \Omega$, to get

$$\begin{aligned}(\widetilde{\pi}P)(y) &= \sum_{t=0}^{\infty} \sum_{x \in \Omega} \mathbf{P}_z\{X_t = x, X_{t+1} = y, \tau_z^+ \geq t+1\} \\ &= \sum_{t=0}^{\infty} \mathbf{P}_z\{X_{t+1} = y, \tau_z^+ \geq t+1\} \\ &= \sum_{t=1}^{\infty} \mathbf{P}_z\{X_t = y, \tau_z^+ \geq t\},\end{aligned}$$

using the fact that the sum of the probabilities of $\{X_t = x\}$ over all $x \in \Omega$ will just equal 1. We notice that this final summation is very similar to our original expression for $\widetilde{\pi}(y)$; in particular,

$$(\widetilde{\pi}P)(y) = \widetilde{\pi}(y) - \mathbf{P}_z\{X_0 = y, \tau_z^+ > 0\} + \sum_{t=1}^{\infty} \mathbf{P}_z\{X_t = y, \tau_z^+ = t\}.$$

But this final term is just accounting for all the occurrences of $X_{\tau_z^+} = y$, and so it sums up to $\mathbf{P}_z\{X_{\tau_z^+} = y\}$, which is equal to 1 when $y = z$ and 0 otherwise (since the Markov chain at its return time should be back at its starting state). Similarly, $\mathbf{P}_z\{X_0 = y, \tau_z^+ > 0\}$ is equal to 1 when $y = z$ and 0 otherwise, since $z = X_0$ and $\tau_z^+ > 0$ are always true by definition. Thus, these two terms are always equal, so they cancel out, leaving us with $(\widetilde{\pi}P)(y) = \widetilde{\pi}(y)\ \forall y \in \Omega$, or

$$\widetilde{\pi}P = \widetilde{\pi}.$$

This proves, that $\widetilde{\pi}$ is stationary, so to make it a stationary distribution, we must divide each element by the sum of the elements, which is equal to

$$\sum_{x \in \Omega} \widetilde{\pi}(x) = \sum_{x \in \Omega} \mathbf{E}_z(\text{number of visits to } x \text{ before returning to } z) = \mathbf{E}_z(\tau_z^+),$$

as the return time for any chain is equal to the total number of states it visits before returning to its start.

Thus, we define

$$\pi(x) = \frac{\widetilde{\pi}(x)}{\mathbf{E}_z(\tau_z^+)},$$

which exists since $\tau_z^+ > 0$ by definition, and we will get a stationary distribution (a stationary vector multiplied by a scalar is still stationary). So by Corollary 2.3, this π is the only such stationary distribution. As such, for any choice of $z \in \Omega$, we will get the same stationary distribution π, so

$$\pi(x) = \frac{\widetilde{\pi}(x)}{\mathbf{E}_x(\tau_x^+)}\ \forall x \in \Omega.$$

Note that choosing $z = x$ also changes the definition of $\widetilde{\pi}$, so that $\widetilde{\pi}(x)$ is now the expected number of visits to x before returning to x, which is exactly 1, for the one time the chain hits x upon returning. Thus,

$$\pi(x) = \frac{1}{\mathbf{E}_x(\tau_x^+)}\ \forall x \in \Omega$$

is a unique stationary distribution for P. □

4. Convergence Theorem

We have now shown that all irreducible Markov chains have a unique stationary distribution π. However, in order to ensure that any distribution over such a chain will converge to π, we require one more condition, called aperiodicity.

Definition 4.1. Let $\mathcal{T}(x) = \{t \geq 1 : P^t(x,x) > 0\}$ be the set of all time steps for which a Markov chain can start and end in a state x. Then the **period** of x is $\gcd \mathcal{T}(x)$.

Lemma 4.2. *If P is irreducible, then the period of all states is equal, or*

$$\gcd \mathcal{T}(x) = \gcd \mathcal{T}(y) \ \forall x, y \in \Omega.$$

Proof. Fix states x and y. Since P is irreducible, $\exists r, l \geq 0$ such that $P^r(x,y) > 0$ and $P^l(y,x) > 0$ (Fig. 2).

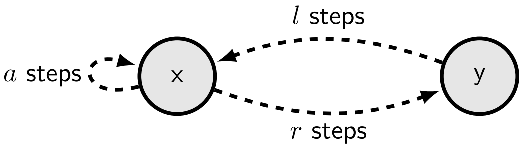

Figure 2. Since P is irreducible, we can get from x to y in r steps, from y to x in l steps, and from x to itself in $a \in \mathcal{T}(x)$ steps.

Let $m = r + l$. Then $m \in \mathcal{T}(x)$, since we can get from x to y in r steps and then from y back to x in l steps, adding up to $r + l = m$ steps. Similarly, $m \in \mathcal{T}(y)$, going from y to x and back to y. If $a \in \mathcal{T}(x)$, there exists a path from x to itself in a steps (Fig. 2), so then $a + m \in \mathcal{T}(y)$ by going from y to x, from x to itself, and from x back to y, totalling $l + a + r = a + m$ steps. Thus, $a \in \mathcal{T}(y) - m \ \forall a \in \mathcal{T}(x)$, where $\mathcal{T}(y) - m = \{n - m \mid n \in \mathcal{T}(y)\}$, so

$$\mathcal{T}(x) \subset \mathcal{T}(y) - m.$$

Take any $n \in \mathcal{T}(y)$, so $\gcd \mathcal{T}(y) \mid n$ by the definition of gcd . Thus, $m \in \mathcal{T}(y)$ implies $\gcd \mathcal{T}(y) \mid m$ as well. This means $\gcd \mathcal{T}(y) \mid n - m \ \forall n \in \mathcal{T}(y)$, or equivalently,

$$\gcd \mathcal{T}(y) \mid a \ \forall a \in \mathcal{T}(y) - m.$$

And we showed that $\mathcal{T}(x) \subset \mathcal{T}(y) - m$, so this also gives us $\gcd \mathcal{T}(y) \mid a \ \forall a \in \mathcal{T}(x)$. So $\gcd \mathcal{T}(y)$ is a common divisor of $\mathcal{T}(x)$, which implies, by the definition of gcd, that

$$\gcd \mathcal{T}(y) \mid \gcd \mathcal{T}(x).$$

By a completely parallel argument, switching around x and y, we also get that $\gcd \mathcal{T}(y) \mid \gcd \mathcal{T}(x)$. Therefore, $\gcd \mathcal{T}(x) = \gcd \mathcal{T}(y) \ \forall x, y \in \Omega$. □

This shows that an irreducible Markov chain has a period common to all of its states, which we then call the period of the chain.

Definition 4.3. An irreducible Markov chain is called **aperiodic** if its period is equal to 1, or equivalently, $\gcd \mathcal{T}(x) = 1 \ \forall x \in \Omega$.

Before being able to prove the Convergence Theorem, we need one result concerning aperiodic chains, and a number-theoretic lemma to prove it.

Lemma 4.4. *If $S \subset \mathbb{Z}^+ \cup \{0\}$ is closed under addition $(a + b \in S\ \forall a, b \in S)$ and $\gcd S = 1$, then there exists M such that $a \in S\ \forall a \geq M$.*

Proof. We begin by showing that there exists a finite subset $T \subset S$ such that $\gcd T = 1$. Let $S_0 = \{a_0\}$, for any $a_0 \in S$. Either $\gcd S_0 = 1$, in which case we let $T = S_0$ and we are done, or there exists $a_1 \in S$ for which $\gcd(S_0 \cup \{a_1\}) < \gcd S_0$, since otherwise we would have $\gcd S = \gcd S_0 \neq 1$. So we let $S_1 = S_0 \cup \{a_1\}$, and then $\gcd S_1 < \gcd S_0$. We continue this process of finding $a_i \in S$ such that if $S_i = \gcd(S_{i-1} \cup \{a_i\})$, then $\gcd S_i < \gcd S_{i-1}$, creating a sequence of finite sets S_i whose gcd decreases until eventually $\gcd S_i = 1$, at which point we let $T = S_i$ and we are done. We know this will occur at some point, since any strictly decreasing sequence of positive integers must hit 1 in a finite number of steps.

Since $\gcd T = 1$, there exists a linear combination of the elements in T that evaluates to 1, so if $T = \{t_1, \ldots, t_n\}$, then there are $c_1, \ldots, c_n \in \mathbb{Z}$ such that

$$c_1 t_1 + \cdots + c_n t_n = 1.$$

Without loss of generality, we can say that all of $c_1, \ldots, c_k \geq 0$ and $c_{k+1}, \ldots, c_n < 0$, for some $1 \leq k \leq n$, so we can move all the negative terms to the other side to get

$$c_1 t_1 + \cdots + c_k t_k = 1 + c_{k+1} t_{k+1} + \cdots + c_n t_n.$$

Since S is closed under addition, and all $t_i \in T \subset S$, the non-negative linear combinations on the left side and right side of this equation are also in S, so if we let these be equal to p and q respectively, we get $p, q \in S$ such that $p = 1 + q$. It is possible that there are no terms other than 1 on the right side, but this just means $p = 1 \in S$, so since S is closed under addition, we can let $M = 1$ and we are done. Otherwise we have $p = 1 + q$ for $p, q \in S$.

Let $M = q(q-1)$. Then for any $a \geq M$, we have $a = kq + r$ for $0 \leq r < q$ and $k \geq q - 1$, by the remainder theorem. Thus $r \leq q - 1 \leq k$, so $k - r \geq 0$, and we can express a as a non-negative linear combination of p and q:

$$a = kq + r = (k - r)q + r(q + 1) = rp + (k - r)q.$$

Therefore, $a \in S\ \forall a \geq M$, as desired. □

Lemma 4.5. *If P is aperiodic and irreducible, then there exists $r \geq 0$ such that*

$$P^r(x, y) > 0\ \forall x, y \in \Omega.$$

Proof. For any $x \in \Omega$, we see that $\mathcal{T}(x)$ is a set of non-negative integers closed under addition, since if $a, b \in \mathcal{T}(x)$, we have $P^a(x, x) > 0$ and $P^b(x, x) > 0$, so $P^{a+b}(x, x) \geq P^a(x, x)P^b(x, x) > 0$ and thus $a + b \in \mathcal{T}$. And since P is aperiodic, $\gcd \mathcal{T}(x) = 1$, which means we can apply Lemma 4.4 to get that there is an M_x such that $P^t(x, x) > 0\ \forall t \geq M_x$. If we then let $M = \max\{M_x : x \in \Omega\}$, we have

$$P^t(x, x) > 0\ \forall t \geq M\ \forall x \in \Omega.$$

Since P is irreducible, there exists a t for any $x, y \in \Omega$ such that $P^t(x, y) > 0$. We let $t_0 = \max\{\text{any one } t \text{ satisfying } P^t(x, y) > 0 : x, y \in \Omega\}$ be the maximum of all of these, and finally let $r = t_0 + M$. Then for any states $x, y \in \Omega$, there exists a $t \leq t_0$ such that $P^t(x, y) > 0$, and then $r - t = M + (t_0 - t) \geq M$, which means $P^{r-t}(x, x) > 0$, so then we get

$$P^r(x, y) \geq P^{r-t}(x, x)P^t(x, y) > 0.$$

□

To prove anything about the convergence of a distribution, we need to first define some measure of distance between distributions.

Definition 4.6. The **total variation** between two distributions μ and ν is defined as

$$||\mu - \nu||_{TV} = \max_{A \subset \Omega} |\mu(A) - \nu(A)|,$$

where $\mu(A) = \sum_{x \in A} \mu(x)$.

Proposition 4.7. $||\mu - \nu||_{TV} = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|$.

Proof. Let $B = \{x \in \Omega : \mu(x) \geq \nu(x)\}$ be the set of states for which $\mu(x) - \nu(x) \geq 0$, so its complement is $B^c = \Omega \setminus B = \{x \in \Omega : \mu(x) < \nu(x)\}$, which is the set of states for which $\mu(x) - \nu(x) < 0$.

Let $A \subset \Omega$ be any set of states. Now A is the disjoint union of $A \cap B$ and $A \cap B^c$, and any $x \in A \cap B^c$ is in B^c and thus $\mu(x) - \nu(x) < 0$, so

$$\mu(A) - \nu(A) = \mu(A \cap B) - \nu(A \cap B) + \mu(A \cap B^c) - \nu(A \cap B^c) \leq \mu(A \cap B) - \nu(A \cap B).$$

And B is the disjoint union of $A \cap B$ and $B \setminus A$, and any $x \in B \setminus A$ is in B and thus $\mu(x) - \nu(x) \geq 0$, so

$$\mu(A \cap B) - \nu(A \cap B) \leq \mu(A \cap B) - \nu(A \cap B) + \mu(B \setminus A) - \nu(B \setminus A) = \mu(B) - \nu(B).$$

Putting these together, we get

$$\mu(A) - \nu(A) \leq \mu(B) - \nu(B),$$

and by symmetric logic with $A \cap B^c$ as an intermediary, we get

$$\mu(B^c) - \nu(B^c) \leq \mu(A) - \nu(A).$$

Now $\mu(B) + \mu(B^c) = \nu(B) + \nu(B^c) = 1$ implies $\mu(B^c) - \nu(B^c) = -(\mu(B) - \nu(B))$, so

$$-(\mu(B) - \nu(B)) \leq \mu(A) - \nu(A) \leq \mu(B) - \nu(B),$$

or $|\mu(A) - \nu(A)| \leq \mu(B) - \nu(B)$. Thus, $|\mu(A) - \nu(A)|$ is bounded by $\mu(B) - \nu(B)$, and we can let $A = B$ to attain this bound, since $|\mu(B) - \nu(B)| = \mu(B) - \nu(B)$ from the fact that $\mu(B) - \nu(B) \geq 0 \ \forall x \in B$. Thus,

$$\begin{aligned}
||\mu - \nu||_{TV} &= \max_{A \subset \Omega} |\mu(A) - \nu(A)| \\
&= \mu(B) - \nu(B) \\
&= \frac{1}{2}\left[\Big(\mu(B) - \nu(B)\Big) + \Big(\nu(B^c) - \mu(B^c)\Big)\right] \\
&= \frac{1}{2}\left(\sum_{x \in B} |\mu(x) - \nu(x)| + \sum_{x \in B^c} |\mu(x) - \nu(x)|\right) \\
&= \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|
\end{aligned}$$

□

Remark 4.8. This proof also gives us $||\mu - \nu||_{TV} = \mu(B) - \nu(B)$. Since $\mu(B) \leq 1$ and $\nu(B) \geq 0$, this tells us $||\mu - \nu||_{TV} \leq 1$, for any distributions μ and ν.

Now we are finally ready to prove the main result of this paper, which tells us that an irreducible, aperiodic Markov chain will converge at an exponential rate to a stationary distribution over time.

Theorem 4.9 (Convergence Theorem)**.** *If P is irreducible and aperiodic, with stationary π, then there exist constants $0 < \alpha < 1$ and $C > 0$ such that*

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

Proof. Since P is irreducible and aperiodic, Lemma 4.5 tells us there exists an $r \geq 0$ satisfying $P^r(x,y) > 0 \ \forall x,y \in \Omega$.

Let $\Pi = \begin{pmatrix} \pi \\ \vdots \\ \pi \end{pmatrix}$ be the $|\Omega| \times |\Omega|$ matrix all of whose rows are the stationary distribution π, making Π a stochastic matrix.

Since $P^r(x,y) > 0 \ \forall x,y \in \Omega$ and $\pi(y) > 0 \ \forall y \in \Omega$, by Theorem 3.3, then

$$\delta' = \min_{x,y\in\Omega} \frac{P^r(x,y)}{\pi(y)} > 0$$

satisfies $P^r(x,y) \geq \delta'\pi(y) \ \forall x,y \in \Omega$. So $\delta = \min\{\delta', \frac{1}{2}\}$ also satisfies this property, as well as $0 < \delta < 1$, so setting $\theta = 1 - \delta$, we get $0 < \theta < 1$.

Now define

$$Q = \frac{P^r - (1-\theta)\Pi}{\theta}.$$

Every element of Q is non-negative, since $P^r(x,y) - (1-\theta)\Pi(x,y) = P^r(x,y) - \delta\pi(y) \geq 0$, by the definition of δ, and, because P^r and Π are stochastic, each row of Q sums to $\frac{1-(1-\theta)}{\theta} = 1$, making Q a stochastic matrix.

For any $n \geq 0$, Q^n is stochastic, from Q being stochastic, so we get $Q^n\Pi = \Pi$, since

$$(Q^n\Pi)(x,y) = \sum_{z\in\Omega} Q^n(x,z)\Pi(z,y) = \pi(y)\sum_{z\in\Omega} Q^n(x,z) = \pi(y) = \Pi(x,y).$$

And since $\pi P = \pi$,

$$\Pi P = \begin{pmatrix} \pi \\ \vdots \\ \pi \end{pmatrix} P = \begin{pmatrix} \pi P \\ \vdots \\ \pi P \end{pmatrix} = \begin{pmatrix} \pi \\ \vdots \\ \pi \end{pmatrix} = \Pi,$$

so $\Pi P^n = (\Pi P)P^{n-1} = \Pi P^{n-1} = \cdots = \Pi P = \Pi$.

Using these identities, we will now prove inductively that

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

The base case $k = 1$ is true, since $P^r = (1-\theta)\Pi + \theta Q$ from the definition of Q. Now assume it is true for $k = n$, so $P^{rn} = (1-\theta^n)\Pi + \theta^n Q^n$. Then

$$\begin{aligned} P^{r(n+1)} &= P^{rn}P^r \\ &= [(1-\theta^n)\Pi + \theta^n Q^n]P^r \\ &= (1-\theta^n)\Pi P^r + \theta^n Q^n P^r \\ &= (1-\theta^n)\Pi + \theta^n Q^n[(1-\theta)\Pi + \theta Q] \\ &= (1-\theta^n)\Pi + \theta^n(1-\theta)Q^n\Pi + \theta^{n+1}Q^{n+1} \\ &= (1-\theta^n)\Pi + (\theta^n - \theta^{n+1})\Pi + \theta^{n+1}Q^{n+1} \\ &= (1-\theta^{n+1})\Pi + \theta^{n+1}Q^{n+1}, \end{aligned}$$

proving inductively that $P^{rk} = (1-\theta^k)\Pi + \theta^k Q^k \ \forall k \geq 1$. Multiplying both sides by P^j for $j \geq 0$,

$$P^{rk+j} = (1-\theta^k)\Pi P^j + \theta^k Q^k P^j = (1-\theta^k)\Pi + \theta^k Q^k P^j = \Pi + \theta^k(Q^k P^j - \Pi),$$

or $P^{rk+j} - \Pi = \theta^k(Q^k P^j - \Pi)$. Looking at any row x of both sides of this equation, we have $(P^{rk+j} - \Pi)(x,\cdot) = \theta^k(Q^k P^j - \Pi)(x,\cdot)$. Thus,

$$\begin{aligned} ||P^{rk+j}(x,\cdot) - \pi||_{TV} &= ||P^{rk+j}(x,\cdot) - \Pi(x,\cdot)||_{TV} \\ &= \frac{1}{2}\sum_{y\in\Omega} |(P^{rk+j} - \Pi)(x,y)| \\ &= \theta^k \frac{1}{2}\sum_{y\in\Omega} |(Q^k P^j - \Pi)(x,y)| \\ &= \theta^k ||Q^k P^j(x,\cdot) - \Pi(x,\cdot)||_{TV} \\ &\leq \theta^k, \end{aligned}$$

since Q, P, and Π being stochastic implies $Q^k P^j(x,\cdot)$ and $\Pi(x,\cdot)$ are distributions, thus their total variation is bounded by 1 from Remark 4.8.

Now we can define the constants $\alpha = \sqrt[r]{\theta}$ and $C = \alpha^{-r}$. Let $t \geq 0$, and define k and j from the division theorem by r, so $t = rk + j$ with $0 \leq j < r$. Finally, since $0 < \theta < 1$ and $0 < \alpha < 1$,

$$\begin{aligned} ||P^t(x,\cdot) - \pi||_{TV} &= ||P^{rk+j}(x,\cdot) - \pi||_{TV} \\ &\leq \theta^k \\ &= \alpha^{rk} \\ &\leq \alpha^{rk}\alpha^{j-r} \\ &= \alpha^{-r}\alpha^{rk+j} \\ &= C\alpha^t. \end{aligned}$$

This is all true for any choice of $x \in \Omega$; therefore, for any $t \geq 0$,

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

□

Acknowledgments. I am deeply grateful to my mentor, Nat Mayer, for devoting his time to learning this subject with me and helping me through the understanding and writing of this paper. I must also thank László Babai for teaching a wonderful and instructive course which gave me much insight into the tools needed for this paper, as well as Peter May for reviewing it and organizing the University of Chicago REU.

References

[1] David A. Levin, Yuval Peres and Elizabeth L. Wilmer. Markov Chains and Mixing Times. http://pages.uoregon.edu/dlevin/MARKOV/markovmixing.pdf.