Markov Chain Theory for Undergraduates

Saad Moro

2019

Acknowledgements

I would like to thank my Honors thesis advisor, Professor James Bernhard, for his help in the writing and editing process, as well as for initially giving me the idea to study the geometry of Markov chains more than a year ago. I would also like to thank Assistant Professor Courtney Thatcher for her support as my reader, and for supervising me while I performed summer research in 2018 on Markov chain theory and its geometric interpretation, without which I would never have been able to write this thesis.

Goals of the text

In this short introduction to Markov chain theory and its applications, we will have a few specific goals in mind. First, I hope to provide a largely self-contained explanation of Markov chain theory. It is expected that the reader be somewhat familiar with linear algebra, and in some of the proofs calculus will be used. Second, I hope that the reader will be able to draw pictures of finite state Markov chains with under 3 states- one for each of our beautiful three dimensions which we are privileged to see. The reader will be able to take a Markov chain and draw how it changes from step to step, and when possible, draw the way in which it converges. Third, this text will promote inter-text readability. We will be using a specific interpretation of Markov chain theory throughout this text that will come naturally to students who have taken linear algebra, however, this interpretation is not the orthodox one in the field of Markov chain theory. There are key differences between most books on Markov chain theory and this pamphlet, so we will highlight those when we come across them.

The goal of this text is not to provide an absolutely rigorous exposition of Markov chain theory in its most formal form, as I feel like that is not conducive to learning. Instead, we will focus more on gaining the intuition on how to understand a modelling problem, transform it into a Markov chain, and then study that model to produce meaningful results about the problem. Some proofs will be shown, in areas where I feel it is worthwhile to highlight the logic behind common results. Everywhere else proofs will be omitted, since they would just be reproductions of standard linear algebra proofs, which can be found in any textbook.

IMPORTANT NOTE: All vectors in this text will be column vectors, and all matrices will be column stochastic (definition of this to come). This is atypical of a text on Markov chains. Usually, Markov chain theory is presented from the perspective of row vectors and row stochastic matrices (should matrices even appear), which is the opposite of the way in which linear algebra is typically taught (column-wise). Since the goal of this text is to provide a pathway from a first course in linear algebra to Markov chain theory, conventions from a first course in linear algebra will be retained.

Chapter 1: Markov Chains and Where to Find Them

In this chapter, we will explore first the general idea behind Markov chain theory, which may sound rather abstract at first, and then gain an intuitive understanding of how Markov chains arise in modelling through looking at examples.

The General Idea

A Markov chain is a probabilistic model that describes a sequence of events, in which the probability of each event depends exclusively on the state attained in the previous step in the sequence.

Before we begin, we will need to introduce a few concepts from probability theory, which will be lightly made reference to later.

Definition 1.1: An **experiment** is some repeatable process, which produces a definite result every time it is done. An individual result of an experiment is called an **outcome.** An **event** is a set of outcomes. The set S of all possible outcomes is called the **sample space**. Importantly for us, a **random variable** is a function from the sample space to the reals, $X : S \to \mathbb{R}$. When we write $P(X = x)$, we mean the probability of the random variable X taking on the value x,, which will result in a value between 0 and 1. The sum of the probabilities of all outcomes is equal to 1. A function with associates every possible outcome of an experiment with its associated probability of occurrence is called a **probability distribution**. We will be dealing with discrete probability distributions, so we can create tables associating every possible value of a random variable with its probability. If we were to do so, the sum of the probabilities in the table would be equal to 1.

Definition 1.2: Conditional probability concerns the probability of some event happening, given that some other event has already occurred. When we write $P(X = x|Y = y)$, we mean the probability that the random variable X has the value x, *given that* the random variable Y attained the value y. With this definition, we may formally define a Markov chain:

Definition 1.3: A sequence of random variables $X_0, X_1, ...$ with values in a finite nonempty

set S is said to be a **Markov chain** when it satisfies the **Markov condition**:

$$P(X_i = s_{l_i} | X_0 = s_{l_0}, X_1 = s_{l_1}, ..., X_{i-1} = s_{l_{i-1}}) = P(X_i = s_{l_i} = X_{i-1} = s_{l_{i-1}})$$

for all $s_{l_0}, s_{l_1}, ..., s_{l_i} \in S$.

In other words, the probability of the current random variable in the sequence having a given value, given the values attained by every random variable before it in the sequence, is equal to the probability of the current random variable in the sequence having a given value given only the value of the random variable that came directly before it in the sequence.

Definition 1.4: With reference to a Markov chain, every value that a random variable in the sequence can attain is called a **state**.

In this text, we consider only **finite-state Markov chains**, i.e. Markov chains in which there is a finite number of values the sequence's random variables can attain. When we say that a Markov chain is "in" some state, we mean that in reference to some $k-th$ moment in **discrete time** (we will consider time to consist of discrete moments), the $k-th$ random variable in the sequence is equal to some state.

Definition 1.5: By the Markov property defined above, since the probability of each event occurring in the $k-th$ moment in time is conditional only on the event that occured in the $(k-1)-th$ moment in time, every state that a moment of time might occupy has with it a set of **transition probabilities**, a probability distribution (set of probabilities whose sum is 1) detailing the probability that the next step in the sequence will attain every possible state. For a Markov chain with n states, we can organize these transition probabilities into an nxn **transition matrix**, where the i, j-th entry is the probability of transitioning from state j to state i.

Definition 1.6: When we create a transition matrix for a Markov chain, the sum of the entries in each column will equal 1. Any matrix which has the sum of the entries in each column equal 1, with each entry greater than or equal to 0, is called a (column)- **stochastic matrix**. In orthodox

Markov chain theory, row stochastic matrices may be used, which as the name implies, have rows whose components sum to 1 and are all greater than or equal to 0.

Three common interpretations of Markov chain theory

The way in which we study Markov chain theory depends largely on the way in which we choose to define a Markov chain. There are three main interpretations of the definition:

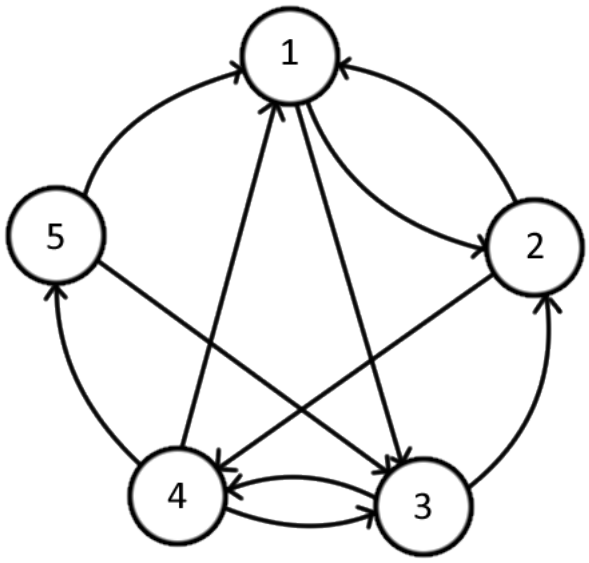
- Markov chains as a finite state automaton consisting of states with associated transition probabilities. This interpretation often involves the creation of graphs describing the different states a Markov chain can attain, and how they relate to each other.
- Markov chains as a collection of random variables satisfying the Markov property of memorylessness.
- Markov chains as linear transformations applied to distribution points.

It is this third lens through which we will study Markov chains.

I believe that it is worthwhile to begin with an example, which will show a situation in which Markov chains can be used to study. From there, we will examine the important structural elements of that example, and from there construct Markov chain theory.

Example 1.7: One avenue through which we interact with Markov chains on a daily basis is the use of internet search engines. At the heart of the Google search engine is an algorithm called PageRank, which organizes search results in order from what is believed to be the most likely webpage for an internet user to want to see, to the least likely. There are multiple components to PageRank, and we will only examine the portion that is relevant to Markov chain theory here, so this is by no means a complete or thorough exposition on the algorithm.

Suppose we have some microcosm of the internet containing only 5 websites, labelled 1 through 5. We can draw a graph consisting of 5 vertices in order to represent our miniature "internet," and whenever one website links to another, we can draw a directed edge between them. Let this relationship between websites and links between them be represented in the graph below:



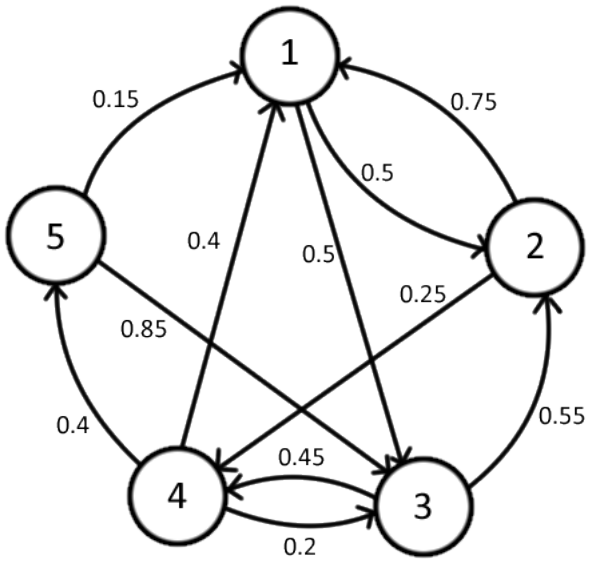
Suppose that we are some memoryless internet surfer. When we are on a website, we locate links from one website to another, and then randomly click one of them. This way, we transition from one website to another, and never think about where we came from. In the context of our internet microcosm, if we are on website 2, there is a probability that we will click a link to website 1, or click a link to website 4. Once we are at the next website, it doesn't matter that we were just at website 2, so if a link back to website 2 exists we might click back to where we just came from.

Now, suppose that some websites are deemed by PageRank to be of a higher quality (or at least are more likely to be the result that internet users want to see when they make a search) than others, resulting in a higher probability of being clicked than other websites. If a memoryless internet surfer was on website 2, we already saw that they have the ability to click links to transition them to websites 1 and 4. Suppose that website 1 is a higher quality website than website 4, so a memoryless web surfer has a probability of 0.75 of clicking to website 1, and a probability of 0.25 of clicking to website 4, and a probability of 0 of clicking to any other website, since presumably website 2 does not link to them. We can express this in a probability distribution:

Website k	P(clicks to website k from website 2)
1	0.75
2	0
3	0
4	0.25
5	0

Notice that the sum of all the probabilities is 1, and every probability is greater than or equal to 0. This means that *something* will happen in the future.

A probability distribution like the one we wrote for website 2 can be written for all websites. We add some sample probabilities to our graph from earlier for the sake of this example:



Beginning at website 2, it is easy to understand which website a memoryless web surfer may end up at after 1 click, but how could we express the probability that they will find themselves at any given website after 2 clicks? 3 clicks? n clicks? If the web surfer clicks for an infinitely long time, how much of their time can we expect to be spent at any given website? These are questions of Markov chain theory.

In a standard text on Markov chains taught through the perspective of probability theory, we would define X_k to be a random variable whose value is w_{l_k} one of the websites from the

set of websites, $\{w_1, w_2, w_3, w_4, w_5\}$, after clicking k links from one website to another. A second subscript is added to w here, since the first (the l)denotes which website was attained at click k, whereas the second subscript (k) denotes which step the Markov chain is at. Our goal is to find $P(X_i = w_i | X_0 = w_2, X_1 = w_{l_1}, ..., X_{i-1} = w_{l_{i-1}})$, which reads as "the probability that $X_i = w_i$ after i clicks, given that after 0 clicks we were at w_2, after 1 click we were at w_{l_1}, and so on. Essentially, this probability is a record of all previous states attained in the process of randomly clicking websites.

This interpretation is difficult to gain an intuition for, and even more difficult to perform computations with. Instead, we can encode all of the information about the probability of transitioning from one website to another inside of a matrix, which we will call a **transition matrix**, T. The i, j-th entry of T is the probability of transitioning from website (or more generally, **state**) j to website (state) i. Since in order to talk about the probability of being at a given website after k clicks (or more generally, *steps*), we need to start somewhere, we will encode information about our starting information in a vector, called the *initial distribution*. In our situation, we have decided that we will start at website 2 always, but we could also have a starting position that is not definite: we could have a 0.5 probability of starting on website 3 and a 0.5 probability of starting on website 5, etc. The columns of the transition matrix are the probability distributions associated with each website, just as we saw the distribution for website 2 earlier. We will call our transition matrix T, and our initial distribution $\vec{d}$.

$$T = \begin{bmatrix} 0 & 0.75 & 0 & 0.4 & 0.15 \\ 0.5 & 0 & 0.55 & 0 & 0 \\ 0.5 & 0 & 0 & 0.2 & 0.85 \\ 0 & 0.25 & 0.45 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0 \end{bmatrix}, \vec{d} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Again, notice that the sum of the entries in every column of both the transition matrix and the initial distribution is 1. Note also that T is a square matrix. This will be the case for every

Markov chain.

How can we use our transition matrix and initial distribution to determine the probability that a memoryless web surfer who begins on website 2 will find themselves on a given website after some number of clicks? Notice that if we multiply T by $\vec{d}$, we get

$$T\vec{d} = \begin{bmatrix} 0 & 0.75 & 0 & 0.4 & 0.15 \\ 0.5 & 0 & 0.55 & 0 & 0 \\ 0.5 & 0 & 0 & 0.2 & 0.85 \\ 0 & 0.25 & 0.45 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.75 \\ 0 \\ 0 \\ 0.25 \\ 0 \end{bmatrix}.$$

Our result is the same distribution we had earlier when talking about performing a single click from website 2! The key here is the *power* of the transition matrix that we multiply our initial distribution by. By what is known as the **Chapman-Kolmogorov equation(s)**, the k-step transition matrix is T^k, for any $k \geq 1$. (The 0-step transition matrix is indeed T^0, but this is just the identity matrix). We will provide a proof, since the Chapman-Kolmogorov equation is absolutely essential to Markov chain theory. Although the vast majority of this text will deal with Markov chains from a linear algebraic approach, we will prove the Chapman-Kolmogorov equation using probability theory due to the proof's reliance on the Markov property, which is most easily stated with reference to probability theory.

Theorem 1.8: Chapman-Kolmogorov equation: For a Markov chain with transition matrix T, the k-step transition matrix is T^k, for any $k \geq 1$.

Proof: [1] In this proof, we will use the notation that $[T]_{ij}$ is the entry of T corresponding to the $i-th$ row's intersection with the $j-th$ column. We use proof by induction. For the base case, $k = 1$, we merely have T. By construction, the 1st step transition matrix is T^1.

For the inductive step, assume that the Chapman-Kolmogorov equation holds for $k \geq 1$. We will show the case for the $k+1$ step. Also, we will use without proof the law of total probability,

[1] Adapted from the proof contained within Durrett, p123.

that the probability of a given event, $P(X_i = i)$ is equal to the sum of its conditional probabilities, $\sum_{X_l, l<i} P(X = i|X_l)$ Then

$$\begin{aligned}
P(X_{k+1} = i|X_0 = j) &= P(X_{k+1} = i|X_k = 1, X_0 = j)P(X_k = 1|X_0 = j) \\
&\quad + P(X_{k+1} = i|X_k = 2, X_0 = j)P(X_k = 2|X_0 = j) \\
&\quad + \cdots + P(X_{k+1} = i|X_k = n, X_0 = j)P(X_k = n|X_0 = j) \\
&= P(X_{k+1} = i|X_k = 1)P(X_k = 1|X_0 = j) \\
&\quad + P(X_{k+1} = i|X_k = 2)P(X_k = 2|X_0 = j) \\
&\quad + \cdots + P(X_{k+1} = i|X_k = n)P(X_k = n|X_0 = j) \\
&\quad \text{(by the Markov property)} \\
&= [T]_{i,1}[T^k]_{1,j} + \cdots + [T]_{i,n}[T^k]_{n,j} \\
&= (i\text{-th row of T}) \cdot (j\text{-th column of } T^k) \\
&= [T^{k+1}]_{i,j}
\end{aligned}$$

So, the probability of being in state i, given that we started in state j after k steps is $[T^{k+1}]_{i,j}$. Since i, j were arbitrary, then the $k - th$ step transition matrix is T^k. $\square$

Returning to our example, the Chapman-Kolmogorov equation tells us that by taking $T^2\vec{d}$, we get the memoryless web surfer's probabilities of being on each website after two clicks, given that they began clicking links on website 2. We have

$$T^2\vec{d} = \begin{bmatrix} 0.375 & 0.1 & 0.5925 & 0.06 & 0 \\ 0.275 & 0.375 & 0 & 0.31 & 0.5425 \\ 0 & 0.425 & 0.09 & 0.54 & 0.075 \\ 0.35 & 0 & 0.01375 & 0.09 & 0.3825 \\ 0 & 0.1 & 0.18 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.375 \\ 0.425 \\ 0 \\ 0.1 \end{bmatrix}$$

Where the i-th entry of $T^2\vec{d}$ is the probability of being on website i after clicking two links. This makes sense, considering the graph of our internet microcosm. For example, starting at website 2, it is impossible to be at website 4 after clicking two links. Thus, the probability of being at website 4 is 0, and sure enough the 4th component of T^2d is 0. Notice also that in the matrix T^2, the sum of the entries in each column is still 1.

We can now find the probabilities of being on any website after any number of clicks. A natural question at this point may be, what is the long term behavior of randomly clicking? If a memoryless web surfer were to click an infinite number of links, what portion of their time would be spent on each website? We will explore this question further in chapter 3, but we can uncover some evidence by continuing to take powers of T. Given that the web surfer started on website 2, after 4 clicks there is a nonzero probability of being on every website. Moreover, after 4 clicks, every entry of the matrix T^4 is greater than 0. Once T^k is entirely positive once, any further power will also be positive. If we take k to be a very high number, we see that every column looks identical: for example, for $k = 100$, we have

$$T^{100} = \begin{bmatrix} 0.2742 & 0.2742 & 0.2742 & 0.2742 & 0.2742 \\ 0.2624 & 0.2624 & 0.2624 & 0.2624 & 0.2624 \\ 0.2279 & 0.2279 & 0.2279 & 0.2279 & 0.2279 \\ 0.1682 & 0.1682 & 0.1682 & 0.1682 & 0.1682 \\ 0.0673 & 0.0673 & 0.0673 & 0.0673 & 0.0673 \end{bmatrix}$$

Now, if we were to multiply T^{100} by any distribution vector (i.e. all entries greater than or equal to 0, and sum to 1), we would always have approximately the same result. This is true because the columns of a matrix are the images of the basis vectors, and if every column is the same, no matter what linear combination of basis vectors the initial distribution is, we will end up with the same result upon multiplication. So it seems that in the long run, it doesn't matter where we started- we always end up with the same distribution eventually. We will explore the properties behind this in upcoming chapter. Merely taking a matrix to a very high power is not

quite enough to show that we will always acquire the same distribution in a matrix-vector product, so we will acquire something much stronger soon.

The vector that results from multiplying a long-term transition matrix with a distribution vector, when organized in order of greatest to least probability, gives us the expected portions of time a memoryless web surfer would spend on each website. With a little bit of formatting, this ordered probability distribution gives us a search engine's search results.

When modelling with Markov chains, there are a few key steps that we must take in order to perform calculations and acquire meaningful information about a system. These are

1. Identify the *states* of the chain.
2. Identify the probabilities for the system to transition from one state to every other state. Every state has its own transition distribution.
3. Organize these distributions into a *column stochastic* matrix, i.e. a matrix where the entries in each column sum to one, with all entries greater than or equal to 0.

In our previous example, states were individual websites, and we decided on transition probabilities from one state to another by looking at which websites linked to one another, and the quality of each website. We will look at one more example to get help get us in the mindset of taking a modelling problem and creating a Markov chain out of it.

Example 1.9: Randomly Generated Text: One common use of Markov chains in computer science is the creation of a random text generator, in which a user inputs a text, and a program generates a new text that looks similar to to it. Small input texts can produce rather goofy results, however very large inputs can product sentence structure, dialogue, and can even mimic an author's writing style. If we are given a sample text, how could we go about creating a Markov chain to randomly generate new text?

One way is as follows: first, we identify the states of the Markov chain to be every unique word in the sample text. At every instance of the word, we look at what the following word is, and extract transition probabilities from one word to the next by dividing the number of occurrences

where that second word follows the first by the total number of instances of the first word. Thus for each word, we acquire a probability distribution of the transition probabilities from every unique word in the text to every other unique word in the text. Then, if we take a "seed word," a word to start with, we can use random number generation to decide which state to transition to given that some word is the most recently generated word.

Chapter 2: Markov Chains and Geometry

In the previous chapter, we saw what a Markov chain is and how to construct one to model a real world situation. Our next goal is to reframe our understanding of Markov chains so that we may gain a geometric intuition of them, and be able to draw pictures of Markov chains and depict their convergence (we have not provided a definition of convergence yet- this belongs in the next chapter), within our dimensional limits.

The key aspect of the geometric understanding to Markov chain theory is the replacement of the Markov chain's transition matrix, T, with the linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ with which it is uniquely associated. To continue, we will need to develop some new vocabulary, and reframe some old terms.

Definition 2.1: We frame a Markov chain's **states** to be the standard basis vectors $\vec{e_1}, ..., \vec{e_n}$ in $\mathbb{R}^n$. Then linear combinations of the standard basis vectors define probability distributions, provided the scalar coefficients in the linear combination are all greater than or equal to 0, and sum to 1.

Now, it is totally possible to use a basis other than the standard one, and still define a Markov chain's states. Throughout this text, we will only use the standard basis vectors, however this need not be the case. We could even define a Markov chain in a vector space other than $\mathbb{R}^n$ and retain tons of information, however for the sake of ease of understanding, we will stick to $\mathbb{R}^n$ and the standard basis vectors thereof.

Definition 2.2: Instead of the transition matrix with which we are familiar, we will call $T : \mathbb{R}^n \to \mathbb{R}^n$ the **transition transformation**, or more precisely the **transition operator**.

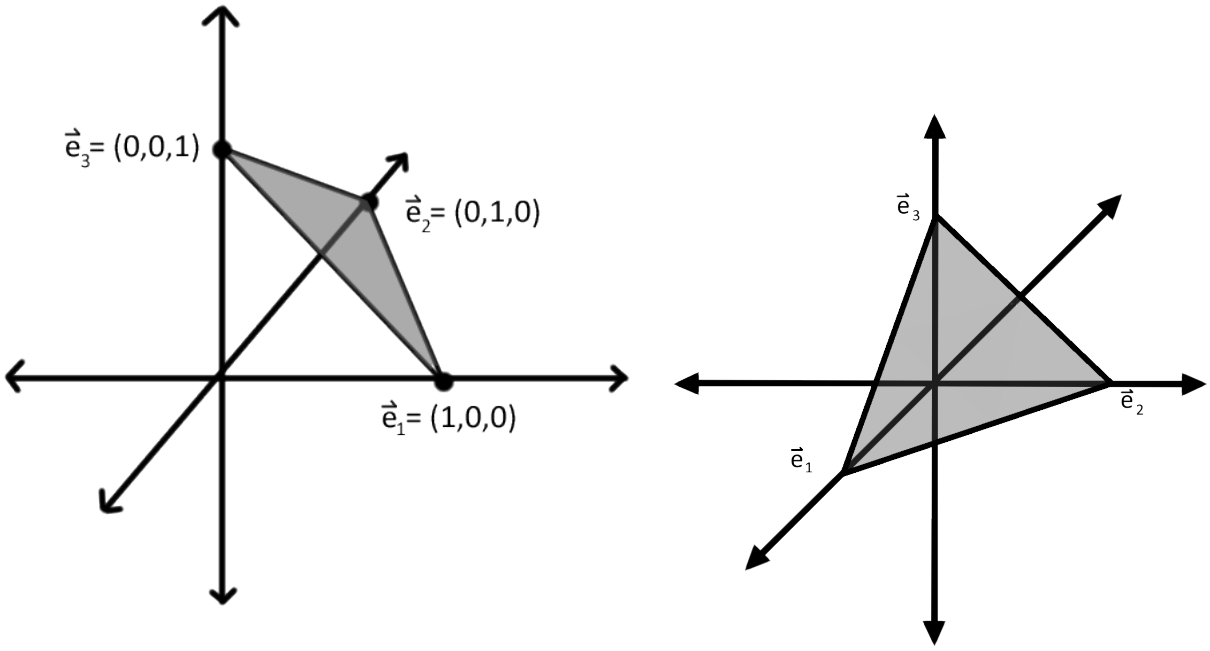
Definition 2.3: A **convex set** is a set such that for any two points in the set, the straight line that connects the two is also contained within the set.

Definition 2.4: The **convex hull** of some points in $\mathbb{R}^n$ is the smallest possible convex set that contains the points in $\mathbb{R}^n$. The standard $(n-1)$-simplex Δ in $\mathbb{R}^n$ is the convex hull of the

first n standard basis vectors $\vec{e_1}, ..., \vec{e_n}$. We can also define the $(n-1)$-simplex in $\mathbb{R}^n$ to be the set $\Delta = \{v_1\vec{e_1} + \cdots + v_n\vec{e_n} | \sum_{i=1}^{n} v_i = 1, v_i \geq 0 \forall i\}$.

Since the sum of all the v_i's is 1, organized in vector format we have a probability distribution vector. Therefore, we recognize the set of all possible distributions in $\mathbb{R}^n$ (distributions with n elements) to be the standard (n-1)-simplex Δ in $\mathbb{R}^n$. .

Example 2.5: The standard 2-simplex Δ in $\mathbb{R}^3$. For any two points on the flat surface of Δ, the line between them is contained entirely in Δ. Note also that any point in Δ can be written as the linear combination of the standard basis vectors $\vec{e_1}, \vec{e_2}, \vec{e_3}$. The point would have the sum of its components sum to 1, just as we would expect a probability distribution to. In the figures below, we show two rotations of the 2-simplex in $\mathbb{R}^3$. The first depicts the two-dimensionality of the simplex, while the second is perhaps an easier to imagine depiction.



One of the most important takeaways from a first course in linear algebra is that in the standard basis, the i-th column of a matrix is the image of the standard basis vector e_i under the matrix's associated linear transformation. In our first example in chapter 1, we saw what happens when we applied a transition matrix to a single distribution vector: we ended up with another distribution

vector. Now that we are talking about linear transformations that have proper domains and images, rather than just matrix-vector products, we can talk about what happens to the *entire* domain of distributions when we apply the linear transformation to them.

For *any* Markov chain, we limit our application of $T : \mathbb{R}^n \to \mathbb{R}^n$ to Δ, the set of distributions in n dimensions, since in the context of a Markov chain there is no meaning in a vector whose sum-of-components is not 1. Since the transition matrix T is column stochastic; that is, its columns consist of entries which are greater than or equal to zero and sum to 1, and whenever we multiply the transition matrix by a distribution, we get another distribution. Then our associated linear transformation may also be called stochastic; that is, $T(\Delta) \subseteq \Delta$, which is an equivalent statement. A more precise statement about a Markov chain's linear transformation would be to define it $T : \Delta \to \Delta$.

Theorem 2.6: Stochastic linear operator characterization: The image of Δ in $\mathbb{R}^n$ under the stochastic linear operator T is a subset of itself, i.e. $T(\Delta) \subseteq \Delta$.

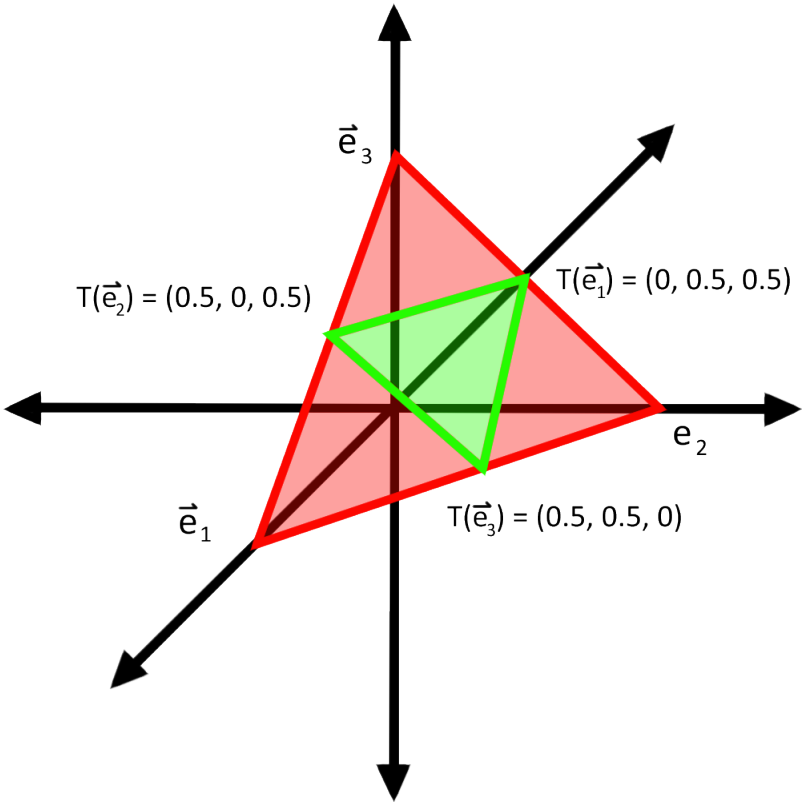
Proof: Any $\vec{x} \in \Delta$ can be written as a linear combination of the standard basis vectors $\vec{e_1}, ..., \vec{e_n}$, whose coefficients are positive and sum to 1. So the entries of $\vec{x}$ are those coefficients, and they sum to 1. Since a linear operator uniquely defines a square matrix M, which in our case is defined to be stochastic, then the sum of the entries in each column is 1. If we take $M\vec{x}$, then by the definition of a matrix-vector product we get another vector, $\vec{y}$, whose components also sum to 1. Therefore $\vec{y}$ can be written as a linear combination of the standard basis vectors, and so $\vec{y} \in \Delta$. Since $\vec{x}$ was arbitrary, it is true that for every $\vec{x}$, $T(\vec{x}) \in \Delta$. Therefore $T(\Delta) \subseteq \Delta$. $\square$

Example 2.7: Consider a Markov chain with transition matrix

$$T = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \end{bmatrix}.$$

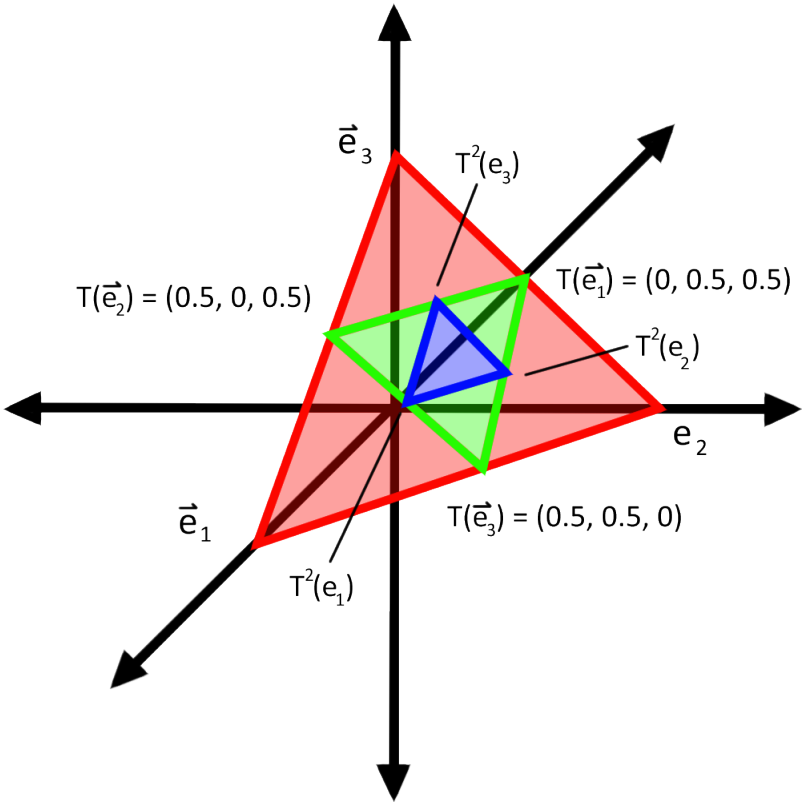
The domain of this Markov chain is the standard 2-simplex in $\mathbb{R}^3$, and its image is contained

entirely within that 2-simplex:



In the above figure, T sends each basis vector to the point in Δ directly opposite from it, whose coordinate is the linear combination of the other two basis vectors. The image of the outer simplex is the inner simplex, because once we have found the image of each basis vector (the corresponding column in T), since every distribution in the original simplex is a linear combination of the basis vectors, then every distribution in $T(\Delta)$ is a linear combination of $T(\vec{e_1})$, $T(\vec{e_2})$, and $T(\vec{e_3})$, which makes the image of Δ under T the inner simplex.

Now, if we were to take $T^2(\Delta)$, *in this case*, we would have $T^2(\Delta) \subset \mathsf{int}(T(\Delta))$, however this is not the case in general. We are only absolutely guaranteed that $T(\Delta) \subseteq \Delta$.



Similarly to the explanation above, the image of $T(\Delta)$ under T, or equivalently the image of Δ under T^2, is the innermost simplex. We find the images of the basis vectors and recognize that all linear combinations thereof are contained within their convex hull. This is because $Hull(T(\Delta)) = Hull(T(\vec{e_1}) + \cdots + T(\vec{e_n}))$.

In this specific case, it looks like each further iteration of the linear transformation T brings the image of Δ under T^k into a smaller and smaller region, and this is indeed the case in the long term!

Chapter 3: Converging Markov Chains

Sometimes, a Markov chain will approach some sort of limit point after we iterate the random process many, many times. We can use this limit point to answer questions such as "in the long run, how long will we spend in a given state?" In this chapter, we will explore the definition of a convergent Markov chain, understand situations from which convergent Markov chains arise, and present the Markov Chain Convergence Theorem.

First, recall the definition of an eigenvalue and eigenvector:

Definition 3.1: Let A be a linear operator (a linear transformation that sends a vector space to itself), or its associated square matrix. A number $\lambda \in \mathbb{R}$ is called an **eigenvalue** of A if there exists some vector $\vec{v}$ such that $A(\vec{v}) = \lambda\vec{v}$. If $\vec{v} \neq 0$, then we call $\vec{v}$ an **eigenvector**.

Definition 3.2: A distribution which does not change when its associated transition matrix is applied is defined as a **stationary state**.

Definition 3.3: Suppose a Markov chain has a stationary distribution $\vec{v}$. If for any initial distribution $\vec{x}$, we have $\lim_{k\to\infty} T^k\vec{x} = \vec{v}$, then we say that our Markov chain is **convergent**, or that it **converges to $\vec{\mathbf{v}}$**.

We can ask questions about convergent Markov chains. What criteria can we place on a Markov chain that dictates when it converges or not? Do all Markov chains with stationary distributions converge?

We begin by continuing our example concerning the PageRank algorithm:

Example 3.4: Before, we had the transition matrix

$$T = \begin{bmatrix} 0 & 0.75 & 0 & 0.4 & 0.15 \\ 0.5 & 0 & 0.55 & 0 & 0 \\ 0.5 & 0 & 0 & 0.2 & 0.85 \\ 0 & 0.25 & 0.45 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0 \end{bmatrix}$$

and we saw that it seemed to converge when we took T^k where k was a very high power. We will take that T here is convergent, however we will be able to provide a rigorous explanation as to why later. Then for any initial distribution $\vec{x}$, $\lim_{k\to\infty} T^k\vec{x} = \vec{v}$, where $\vec{v}$ is the stationary distribution, or the eigenvector of T. Using a computational program such as Wolfram Alpha, we can compute the eigenvalue associated with eigenvalue 1:

$$\vec{v} = \begin{bmatrix} \frac{2731}{670} \\ \frac{1307}{335} \\ \frac{227}{67} \\ \frac{5}{2} \\ 1 \end{bmatrix}$$

However, it is clear that the sum of the components of $\vec{v}$ do not equal 1. Without defying the definition of an eigenvector, we can multiply $\vec{v}$ by 1 over the sum of its components in order to normalize it such that the sum of its components is 1, making it an acceptable stationary distribution.

$$\vec{v} = \begin{bmatrix} \frac{2371}{9960} \\ \frac{1307}{4980} \\ \frac{227}{996} \\ \frac{335}{1992} \\ \frac{67}{996} \end{bmatrix} \approx \begin{bmatrix} 0.2742 \\ 0.2625 \\ 0.2279 \\ 0.1682 \\ 0.0673 \end{bmatrix}$$

So, after an infinite number of clicks, our original memoryless web surfer could expect to spend portions of their time on each website equal to the components in the above vector.

We now will construct the central result about the convergence of Markov chains.

Lemma 3.5: Let A be a square matrix with eigenvalue λ and corresponding eigenvector $\vec{v}$. Then

$$A^n\vec{v} = \lambda^n\vec{v}$$

holds for every positive integer n.

Proof: We use induction. For the base case, let $n = 1$. This statement is true by the definition of an eigvalue and eigenvector:

$$A\vec{v} = \lambda\vec{v}$$

For the inductive step, assume that $A^k\vec{v} = \lambda^k\vec{v}$. Then we just need to show that $A^{k+1}\vec{v} = \lambda^{k+1}\vec{v}$. We have

$$\begin{aligned}
\lambda^{k+1}\vec{v} &= \lambda\lambda^k\vec{v} \\
&= \lambda A^k\vec{v} && \text{inductive hypothesis} \\
&= A^k\lambda\vec{v} && \text{scalar multiplication} \\
&= A^k A\vec{v} && \text{definition of eigenvector} \\
&= A^{k+1}\vec{v}
\end{aligned}$$

Therefore for $n \in \mathbb{N}$, $A^n\vec{v} = \lambda^n\vec{v}$. $\square$

This next theorem is important in obtaining a right eigenvector in the upcoming proof of the Perron-Frobenius theorem, which will give us an understanding of the criteria for converging Markov chains. No proof will be provided since it is outside of the scope of this text, but we will explain why it is applicable to our needs.

Definition 3.6: A define a vector $\vec{v} \in \mathbb{R}^n$ to be **non-negative** if it has no negative components, i.e. $\vec{v} \geq \vec{0}$. A vector $\vec{u} \in \mathbb{R}^n$ may be defined to be a **positive** vector if its components are *strictly* greater than 0, i.e. $\vec{u} > \vec{0}$.

Definition 3.7: We define the **interior** of the set $\Delta \subset \mathbb{R}^n$, written $\mathsf{int}(\Delta)$ to be the set of all points $\vec{v} \in \Delta$ such that the linear combination $\vec{v} = v_1\vec{e_1} + \cdots + v_n\vec{e_n}$ has all $\vec{v_i}$ strictly greater than 0, for all i.

Definition 3.8: We define the **boundary** of the set $\Delta \subset \mathbb{R}^n$, written $\mathsf{bd}(\Delta)$ to be the set of all points $\vec{v} \in \Delta$ such that the linear combination $\vec{v} = v_1\vec{e_1} + \cdots + v_n\vec{e_n}$ has $\sum_{i=1}^{n} v_i = 1$ but with at least one $v_i = 0$.

Definition 3.9: Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a linear operator, and T_M be its associated unique stochastic matrix. T_M is defined to be a **positive matrix** if every entry is strictly greater than 0. T is defined to be a **positive linear operator** is $T(\Delta) \subset \mathsf{int}(\Delta)$. These definitions are equivalent,

because the image of the standard basis vectors under T will always be strictly positive, and $T(\Delta)$ is the convex hull of the image of the standard basis vectors under T.

Theorem 3.10: Brouwer Fixed-Point Theorem Let f be a continuous function that maps a convex, compact set $\mathcal{C}$ to itself. Then there exists some point $x_0 \in \mathcal{C}$ such that $f(x_0) = x_0$.

Proposition 3.11: Properties of Δ: The set Δ is convex and compact.

Proof: In the previous chapter, we defined $\Delta = \{v_1\vec{e_1} + \cdots + v_n\vec{e_n} | \sum_{i=1}^{n}(v_i) = 1, v_i \geq 0 \forall i\}$ to be the $(n-1)$-dimension standard simplex in n dimensions, i.e. the convex hull of the first n standard basis vectors. So by construction, Δ is a convex set.

A set in $\mathbb{R}^n$ is compact when it is closed and bounded. Let $S = \{\vec{x} | \vec{x} \in R^n, ||\vec{x}|| \leq 1\}$ be the sphere of length 1 around the origin. Since any $\vec{v} \in \Delta$ can be written $\vec{v} = v_1\vec{e_1} + \cdots + v_n\vec{e_n}$ with $\sum_{i=1}^{n} v_i = 1$, then $||\vec{v}|| = \sqrt{v_1^2 \cdots + v_n^2} \leq \sqrt{1} = 1$. Since $\vec{v} \in \Delta$ was arbitrary, within R^n we have $\Delta \subset S$. So Δ is bounded. If we were to construct some congergent sequence $\{\vec{v_n}\} \in \Delta$, then any point to which the sequence may convergence is also an element of Δ, i.e. the edges of Δ are contained within Δ as well. Thus Δ is closed. Since Δ is closed and bounded, it is compact. $\square$

The next theorem, the Perron-Frobenius Theorem, gives us that the stationary distribution of a converging Markov chain is unique.

Theorem 3.12: Perron-Frobenius Theorem Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a stochastic linear transformation. If for some positive integer $k \in \mathbb{N}$, $T^k \subset \mathsf{int}(\Delta)$, then

1. T has an eigenvalue of 1, λ_1, and for any other eigenvalue λ_i associated with eigenvector $\vec{x_i}$ with $i \neq 1$, $|\lambda| < 1$.

2. The eigenvalue 1 has geometric and algebraic multiplicity 1.

3. The eigenvector associated with the eigenvalue 1 is positive.

Proof: Although it is not assumed that T is positive from the outset, by the lemma above we can pick some positive integer k such that $T^k(\Delta) \subset \mathsf{int}(\Delta)$. Let $A = T^k$ be chosen for a value of k such that $A(\Delta) \subset \mathsf{int}(\Delta)$. Also, let $Q = \{\vec{u} = \sum_{i=1}^{n} u_i\vec{e_i} | u_i \geq 0 \forall i\}$ be the positive orthant

(generalized positive quadrant). We will make occasional reference to Q throughout this proof. Notice that $\Delta \subset Q$, and we could write any member of Q as a linear combination of vectors in Δ. We could easily use Δ instead, but Q allows us to not have to normalize every vector so that it lies on the surface of Δ.

In the proposition above, we proved that Δ is a set that is both convex and compact. Then if we apply A to Δ, by the Brouwer Fixed Point theorem, we get that there exists some $x_1 \in \Delta$ such that $A(x_1) = x_1$. By the definitions of eigenvalue and eigenvector, x_1 is an eigenvector associated with the eigenvalue 1. This proves the first part of part (1) of the theorem. Moreover, x_1 is strictly positive, since A is assumed to be positive, and $A(\vec{x_1}) \subset \mathsf{int}(\Delta)$. This proves part (3) of the theorem.

Let $\mathbb{1}$ be the vector of length n, with each entry equal to 1. Since A is a stochastic linear transformation, it is associated uniquely with a stochastic matrix, and by the definition of a vector-matrix product, $\mathbb{1}^\top A = \mathbb{1}^\top$ (since the sum of all columns in A is 1, a row vector multiplied by A yields a row vector of 1s). Since $\mathsf{dim}(\mathbb{R}^n) = n$, A has *at most* n eigenvalues, and we just showed that A has at least one eigenvalue, so let the number of eigenvalues be $m \in [1, n]$. We will write our eigenvalues as $\lambda_1, \lambda_2, ..., \lambda_m$ and our eigenvectors $\vec{x_1}, \vec{x_2}, ..., \vec{x_m}$, with each index associating an eigenvalue with an eigenvector. Our next step will be to determine that the geometric and algebraic multiplicity of the eigenvalue 1 is 1.

First, we need to show that a vector $\vec{x}$ which is positive, but with at least 1 zero in its components (i.e. $\vec{x} \geq \vec{0}$ but not $\vec{x} > \vec{0}$), cannot be an eigenvalue of A. Since we have that $A(Q) \subset \mathsf{int}(Q)$, and $\vec{x} \in Q$, then $A\vec{x}$ would have to be an element of $\mathsf{int}(Q)$. Suppose that $\vec{x} \geq \vec{0}$ but not $\vec{x} > \vec{0}$ is an eigenvector associated with the eigenvalue 1. Then by the definition of an eigenvector associated with eigenvalue λ, we have $A\vec{x} = \lambda\vec{x} = \vec{x}$. But $\vec{x} \notin \mathsf{int}(Q)$, and $A\vec{x} \in \mathsf{int}(Q)$, so we have a contradiction. $\vec{x}$ cannot be an eigenvalue.

Next, consider the case where we have two eigenvectors associated with the eigenvalue 1, both strictly positive, $\vec{x}$, and $\vec{w}$, and neither are scalar multiples of one another. Consider $\vec{y} = \vec{x} - \alpha\vec{w}$, where α is a scalar. The vector $\vec{y}$ is an eigenvector associated with the eigenvalue 1, since $A\vec{y} = A(\vec{x} - \alpha\vec{w}) = A\vec{x} - A\alpha\vec{w} = A\vec{x} - \alpha A\vec{w} = \vec{x} - \alpha\vec{w} = \vec{y}$, so $\vec{y}$ satisfies the definition of an eigenvalue.

Now choose α to the maximum possible value such that $\vec{y} \geq \vec{0}$, but $\vec{y} \not> \vec{0}$. Such an α exists since we assumed $\vec{x}$ and $\vec{w}$ not to be scalar multiples of each other, and both $\vec{x}$ and $\vec{w}$ are strictly positive. But above we proved that an eigenvector associated with the eigenvalue 1 cannot be greater than the zero vector and not strictly positive. So we have arrived at a contradiction. Since $\vec{x}$ and $\vec{w}$ could be any strictly positive eigenvectors associated with the eigenvalue 1, but we did previously prove that at least one such eigenvector exists, we have arrived at the conclusion that *no more than one* such eigenvector can exist. Next we will set up a subspace which will allow us to prove that no vectors which are a mixture of positive and negative components can be an eigenvector associated with the eigenvector 1, and also that the absolute value of every other eigenvalue must be less than 1.

We have that $\mathbb{1}^\top$ is a left eigenvector and $\vec{x_1}$ is a right eigenvector associated with the eigenvalue 1. We can use $\mathbb{1}$ to show that $\vec{x_1}$ has geometric and algebraic multiplicity 1. Define the $(n-1)$ dimensional subspace W of $\mathbb{R}^n$ to be $W = \{\vec{x} \in \mathbb{R}^n : \mathbb{1}^\top \vec{x} = 0\}$. Any vector $\vec{x} \in W$ is orthogonal to $\mathbb{1}$, and moreover since every entry of $\mathbb{1}$ is strictly positive, then $\vec{x}$ must have some positive components and some negative components. Therefore no $\vec{x} \in W$ could be an eigenvector associated with the eigenvalue 1. Take some eigenvalue λ_i with $1 < i \leq m$, and its associated eigenvector $\vec{x_i}$. Since $\vec{x_i} \in W$, then the components of $\vec{x_i}$ are a mixture of positive and negative numbers. Consider the two dimensional subspace of $\mathbb{R}^n$, $Y = \mathsf{Span}(\vec{x_1}, \vec{x_i})$. For any $\vec{v} \in Y$, we can write $\vec{v} = \theta_1 \vec{x_1} + \theta_2 \vec{x_i}$, and $A(\vec{v}) = A(\theta_1 \vec{x_1} + \theta_2 \vec{x_i}) = \theta_1 A(\vec{x_1}) + \theta_2 A(\vec{x_i}) = \theta_1 \vec{x_1} + \theta_2 \lambda_i \vec{x_i}$. We can choose choose some x_a such that $A^n(x_a) \in Q$. If $|\lambda_i| > 1$, then $\lim_{n\to\infty} A({}^n x_a)$ would converge to some point in the subspace $\mathsf{Span}(\vec{x_i}) \notin Q$, since eventually $\theta_2 |\lambda_i|^k \vec{x_i} > \theta_1 \vec{x_1}$. So for any eigenvalue $\lambda_2, ..., |\lambda_i| < 1$ where $i \geq 2$. This proves the second half of part (1) of the theorem.

Then if $|\lambda_i| = 1$, consider how we could also write any vector on the unit circle of length 1 around the origin to be $\vec{u} = cos(\theta)\vec{x_1} + sin(\theta)\lambda_i \vec{x_i}$. Let $\vec{u}$ be a vector on the intersection of the unit sphere and $\mathsf{bd}(Q)$, the boundary of Q, such that $\vec{u}$ is also in W. If we let $k = 2$, then by the Pythagorean identity we would have $T^2(\vec{u}) = \vec{u}$, but since we assumed T to be positive, this contradicts the fact that $T^2(\vec{u})$ must be inside the interior of Δ. Therefore it is *not true* that in our situation, an eigenvector can be associated with the eigenvalue 1 and have a mixture of positive

and negative components.

The situation where an eigenvector's components are entirely negative is just a scalar multiple of the case where its components are entirely positive. So, we have shown that an eigenvector of the linear operator A associated with the eigenvalue 1 cannot have its components be a mixture of positive and negative, and it cannot be greater than zero, but not strictly greater than 0. It must be a strictly positive eigenvector, and moreover, there only exists one such strictly positive eigenvector associated with the eigenvalue 1. Therefore, the geometric multiplicity of the eigenvalue 1 must be 1. There exists an identity [2] which states that if an eigenvalue has geometric multiplicity 1, then it has algebraic multiplicity 1 if and only if the left and right eigenvectors associated with the eigenvalue are not orthogonal. Since we have shown $\mathbb{1}$ to be a left eigenvector, and $\vec{x_1}$ to be a right eigenvector. Since both eigenvectors are strictly positive, then $\mathbb{1} \cdot \vec{x_1} > 0$. Therefore $\mathbb{1}$ and $\vec{x_1}$ are not orthogonal, and so the eigenvalue 1 has algebraic multiplicity 1. This proves part (2) of the theorem.

By the lemma above, since $A = T^k$, then the eigenvalues we found belonging to A are the k-th powers of the ones belonging to T. Since $\lambda_1 = 1$, then its associated eigenvalue in T is also 1, and since all other eigenvalues were found to be less than 1, their corresponding eigenvalues in T are also less than 1. Also by the lemma above, the eigenvector remains unchanged despite the raising of power of T. So everything that is true about the eigenvalues and eigenvectors of A remains true for T, and the theorem is proved.

□

Theorem 3.13: Markov Chain Convergence Theorem Let T be a stochastic linear transformation such that for some positive integer k, $T(\Delta) \subset \mathsf{int}(\Delta)$. Let $\vec{z}$ be the eigenvector associated with eigenvalue 1. Then $\lim_{n\to\infty} T^n = Z$ exists, and the columns of Z are all $\vec{z}$. That is, $\lim_{n\to\infty} T^n\vec{x} = \vec{z}$ for any initial distribution $\vec{z}$.

Proof: This theorem was proved when we discussed the case in which another eigenvalue, $|\lambda_i| > 1$ had the linear operator converge to its associated eigenvector. The other case is that when

[2] See Horn & Johnson, p80.

the greater eigenvalue is 1, we converge to the eigenvalue $\vec{x_1}$ as defined in the previous proof. $\square$

To convert this theorem into a matrix-format, let T be a stochastic matrix with $[T^k]_{i,j} > 0$ for all i, j for some positive integer k.

Example 3.14 : This is a continuation of example 2.2 in the previous chapter. For our linear transformation, T, we had

$$T = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \end{bmatrix}.$$

If we take T^2, we have

$$T^2 = \begin{bmatrix} 0.25 & 0.375 & 0.375 \\ 0.375 & 0.25 & 0.375 \\ 0.375 & 0.375 & 0.25 \end{bmatrix}$$

i.e. $T^2(\Delta) \subset \mathsf{int}(\Delta)$ since every column of T^2 is an interior point of Δ. So, the Markov chain convergence theorem applies here. Then there exists some $\vec{v} \in \Delta$ such that $T\vec{v} = \vec{v}$. To find the vector $\vec{v}$, we calculate the eigenvector associated with the eigenvalue $\lambda = 1$. Using any method, we find the eigenvector

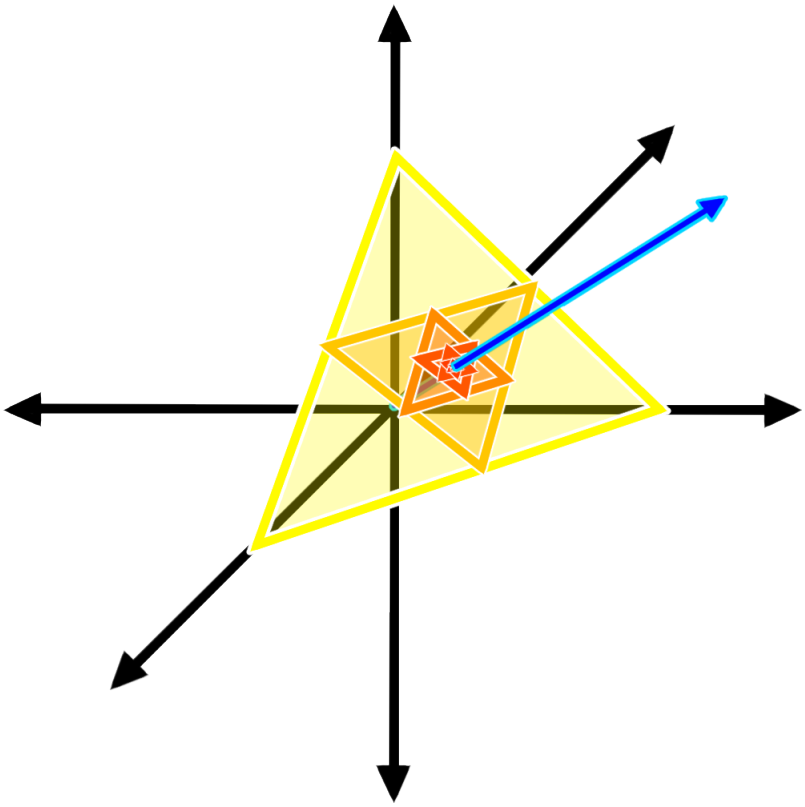
$$\vec{v_1} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Now, while this certainly is an eigenvector that satisfies $T\vec{v_1} = \vec{v_1}$, $\vec{v_1}$ is not quite the answer we are looking for, because its entries do not sum to 1, so it is not a distribution. Fortunately, if we

multiply the eigenvector by a certain scalar (1 / sum of all components of $\vec{v_1}$), we will normalize the eigenvector to have its entries sum to 1, making it a stationary distribution.

$$\frac{1}{\sum_{i=1}^{3} \vec{v}_{1_i}} \vec{v_1} = \frac{1}{1+1+1} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} = \vec{v}$$

Graphically, we see that successive iterations of the linear transformation squeeze the range into a single point, the intersection of the eigenvector associated with eigenvalue 1, and the surface of Δ:



This concludes our discussion of convergent Markov chains. Not all Markov chains converge, however. Our next chapter explores the properties of non-convergent chains in detail, in a way which may also contextualize what we have learned in this chapter.

Chapter 4: Beyond Convergence

In the previous chapter, we focused our attention on converging Markov chains. The geometric view of Markov chains which we first introduced in chapter 2 becomes particularly useful when we begin to consider Markov chains which do *not* converge. We will construct a couple of brief examples, and then look at the theory.

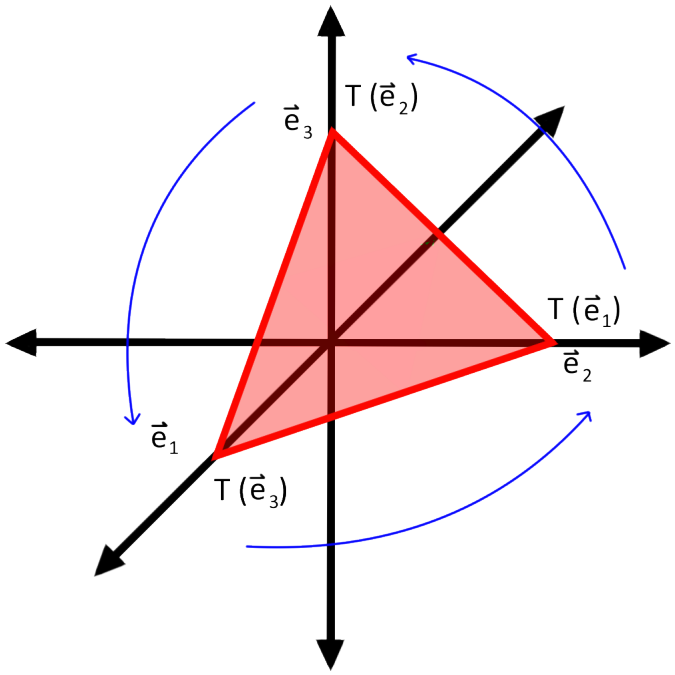
We will provide an example here of one which does not converge. Let

$$T = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

The first few powers of T are

$$T^2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, T^3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, T^4 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Since $T^4 = T$, we see that any further powers of T will give us a matrix whose value cycles through that of T, T^2 or T^3. So in this case, there is no possible point to which $T^k(\Delta)$ could converge to, since in this case for any power k, $T^k(\Delta) = \Delta$. We graphically depict our situation below:



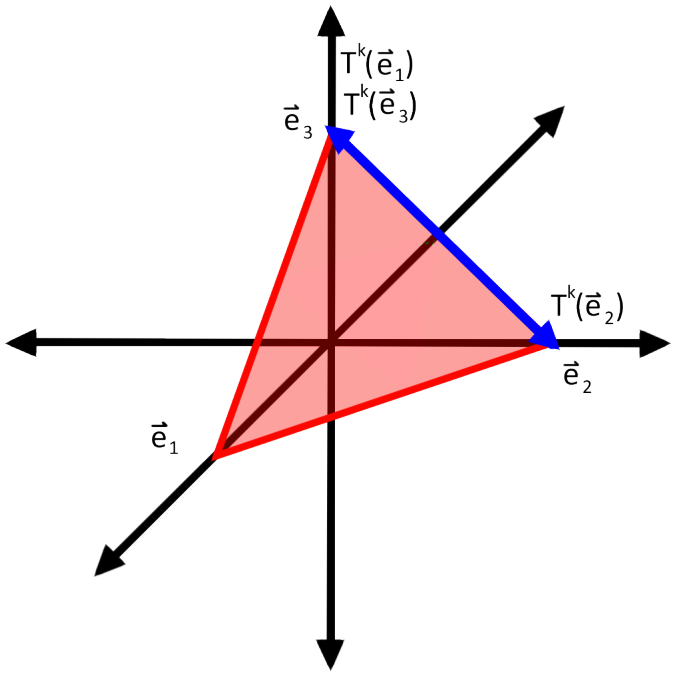
Or, consider a new Markov chain associated with the transition matrix

$$T = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

If we look at the first few powers of T, we find

$$T^2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, T^3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \ldots$$

T clearly does not change at all with respect to each moment in time. Graphically, this looks like the figure below:



T sent $\vec{e_2}$ and $\vec{e_3}$ to themselves, but sent $\vec{e_1}$ to $\vec{e_3}$. Here, the image of Δ under T is radically different than the original shape of Δ, and is totally stationary, invariant under higher powers of T.

What is going on with these two examples, geometrically? If we consider the subset of Δ after each linear transformation, it appears not to change in both examples- although points in the first example might endlessly rotate around the center of Δ, and in the second example the entire simplex is pressed against the edge between $\vec{e_2}$ and $\vec{e_3}$, the subsets seem to be stable. What happens when we consider the set that is the intersection of all subsets of Δ when we take $T^k(\Delta)$ to any power? We will need to create some more vocabulary, and then study our new subjects. The proofs in this section are largely adapted from Bernhard (2012).

Definition 4.1: A set $P \subset \mathbb{R}^n$ is called a **convex polytope** if it is the convex hull (defined in chapter 2) of a finite set of points. So any point inside a convex polytope may be written as the linear combination of those points, whose coefficients sum to 1. If the coefficient in the linear combination of one of the points in the set is 1, then we call that point a **vertex**. Thus a convex

polytope is the convex hull of its vertices.

Lemma 4.2: Convex polytopes are compact.

Proof: Consider the "Properties of Δ" proposition in the previous chapter. The proof therein can be applied to any convex polytope, since it is the hull of a *finite* number of points. $\square$

Theorem 4.3: Let $P_0, P_1, ...$ be a sequence of convex polytopes with $P_0 \supseteq P_1 \supseteq P_2 \supseteq ....$ Then $\bigcap_{k=0}^{\infty} P_k$ is a nonempty set.

Proof: By the above lemma, convex polytopes are compact. This implies that the set $\bigcap_{k=0}^{\infty} P_k$ is compact as well, since the intersection of compact sets is itself compact. Then for every index $k \in \mathbb{N}$, pick some point $\vec{p_k} \in P_k$. Since $\bigcap_{k=0}^{\infty} P_k$ is compact, then the sequence $\{\vec{p_0}, \vec{p_1}, ...\}$ must have a convergent subsequence. Since the intersection of compact sets is compact, and compact sets are closed and bounded, the limit point of a sequence in a compact set must also be in that set. Therefore since the limit of the sequence of points defined in this proof exists, and is in $\bigcap_{k=0}^{\infty} P_k$. Therefore $\bigcap_{k=0}^{\infty} P_k$ is nonempty. $\square$

Definition 4.4: Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a stochastic linear transformation. Then we define the **limit set** L of T to be

$$L = \bigcap_{k=0}^{\infty} T^k(\Delta).$$

Proposition 4.5: L is nonempty.

Proof: Since Δ is by definition a convex polytope, and since $T^k(\Delta)$ is the convex hull of the k-th power of T applied to the standard basis vectors, and since $\Delta \supseteq T(\Delta) \supseteq T^2(\Delta)$ by the stochastic linear operator characterization theorem in chapter 2, we can apply the above theorem to L to say that it is nonempty. $\square$

Theorem 4.6: Transforming the limit set by the transition operator yields the limit set: $T(L) = L$.

Proof: Let $\vec{x}$ be a point in the limit set of T. Then for all $k > 0$, $\vec{x} \in T^k(\Delta)$. Unravelling the sequence that leads to the limit set, we see that $T(\vec{x}) \in T^{k+1}(\Delta) \subset T^k(\Delta)$ for all positive k, so $T(\vec{x}) \in L$ as well. Since this is true for any arbitrary $\vec{x} \in L$, then $T(L) \subset L$. To show equality of two sets, we want to show that they are subsets of one another. So we just need to show that $L \subset T(L)$.

Now, let $\vec{x} \in L$. Then for any $k \geq 1, \vec{x} \in T^k(\Delta) = T(T^{k-1}(\Delta)$. This implies that for for all $k \in \mathbb{N}$, there exists a point $\vec{p_k} \in T^{k-1}(\Delta)$ such that $T(\vec{p_k}) = \vec{x}$. By the compactness of $\Delta \in \mathbb{R}^n$, (see chapter 3), the sequence $\{\vec{p_1}, \vec{p_2}, ...\}$ must have a convergent subsequence that converges to $\vec{p_0} \in \Delta$. By the proof of the nonemptiness of L, then $\vec{p_0} \in L$. By the continuity of T, we have $T(\vec{p_0}) = T(\lim_{n\to\infty} \vec{p_n} = \lim_{n\to\infty}(\vec{p_n}) = \lim_{n\to\infty} \vec{x} = \vec{x}$. So then any arbitrary $\vec{x} \in L$ can be written as $T(\vec{p_0})$ for some point $\vec{p_0} \in L \subset \Delta$. Therefore $T(L) \subset L$.

Since $L \subset T(L)$, and $T(L) \subset L$, then $T(L) = L$. $\square$

Now this is something very interesting! In our discussion of convergent Markov chains, we saw a unique stationary distribution $\vec{v}$ such that $T(\vec{v}) = \vec{v}$. Here, we have an entire limit set, which does not guarantee the existence of a set of stationary points, but in linear algebraic expression looks very similar to the formulation for the stationary distribution we saw previously. On one hand, we have a unique $\vec{v} \in \Delta$ such that $T(\vec{v}) = \vec{v}$. Here, we have $T(L) = L$.

When we look at Markov chains geometrically, we can see that both convergent and non-convergent chains fit into a greater system. In both cases, $L = \bigcap_{k=0}^{\infty} T^k(\Delta)$ exists and is nonempty, however in the convergent case $L = \{\vec{v}\}$, the set of just the unique stationary distribution. In the non convergent case, L merely has a greater number of elements. For the limit sets of convergent and non-convergent Markov chains, we have $T(L) = L$. The geometry of L grants us a greater understanding of the mechanisms underlying both convergent and non-convergent Markov chains.

Conclusion

As we saw in the first chapter of this text, Markov chains are probabilistic models which can be used in a variety of real world situations. We saw how Markov chains are used by the PageRank algorithm used in the Google search engine to determine the ranking of websites which a web surfer would most probably spend their time at, and how a text generator could take a text and induce a Markov chain based on the ordering of words. We introduced these examples through a fairly standard lens of Markov chain theory, but in chapter 2 we introduced a new way to look at Markov chains and the transition matrices and distributions which characterize them. In chapters 3 and 4 we examined the behavior of Markov chains when their transition operators are taken to very high powers, and found that perhaps, convergent and non-convergent Markov chains are not so different after all.

The reader may leave this text with an insight that might not be apparent from the usual probabilistic way of teaching Markov chain theory. By looking at the limit of a transition operator on a set of distributions, we see connections between between the cases of convergent and non-convergent Markov chains, which would otherwise not be transparent. The convergent Markov chain's sought-after initial distribution is but a manifestation of a more general limit set, which exists in common and is nonempty for all Markov chains. Moreover, we developed a geometric intuition of the image of the set of distributions, Δ, under powers of the transition operator T^k. We were able to draw the way in which convergent Markov chains converge in chapter 3, and in chapter 4 we depicted what the limit set of a non-convergent Markov chain looks like.

References

[1] J. Bernhard (2012) The Geometry of Markov Chain Limit Theorems. *Markov Processes and Related Fields* **19**, 99-124.

[2] Robert Horn & Charles Johnson. *Matrix Analysis*. Cambridge University Press, 2013.

[3] Rick Durrett. *Probability Theory for Applications*. Cambridge University Press, 2009.

[4] Sheldon Axler. *Linear Algebra Done Right*. Springer, New York, 2015.

[5] Shlomo Sternberg. *Dynamical Systems*. Dover, 2013.