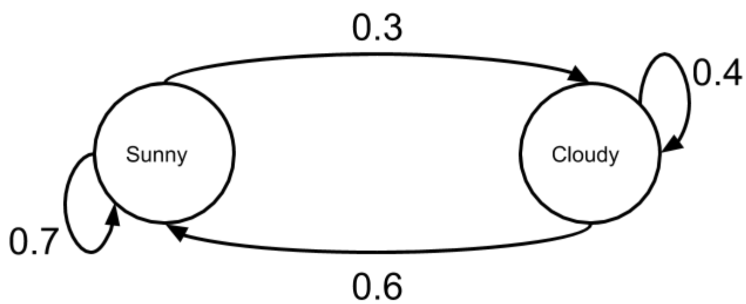
APPM 2360 Project 2
Network Markov Chains

Due: Thursday, November 1, 2018 by 4:59 p.m.
Submit as a PDF to "Assignments" on Canvas

---

1 Introduction

A "Markov Chain" is a common model to describe processes which have multiple states they move between. For example, a day can be cloudy or sunny. If the probability it is cloudy or sunny tomorrow depends only on if it is cloudy or sunny today, we can draw the system so:



We read this diagram by assuming the process moves from state to state with the probabilities given on the arrows. So if it's sunny today, we *know* it's sunny, so we assume we are in the Sunny state with probability $P = 1$ and Cloudy with probability $P = 0$. At the next step (tomorrow), there is a 70%, ie $P = 0.7$, chance we are still in that first state. So we would describe our new position with the probability vector $\langle 0.7, 0.3\rangle$. In this way, the probabilities begin to "flow" from state to state.

We often express Markov chains such as the one pictured above in terms of a transition matrix where each row in the matrix corresponds to the weights on the arrows that flow into that state, and the columns correspond to the weights on the arrows that originate in that state. In the cloudy/sunny case above, for example, we could express the system with the following matrix:

$$\begin{pmatrix} 0.7 & 0.6 \\ 0.3 & 0.4 \end{pmatrix}$$

If we *know* it is sunny today, we express that with the state vector

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Then if we want to know the probabilities of the weather tomorrow, we multiply that vector by the matrix:

$$\begin{pmatrix} 0.7 & 0.6 \\ 0.3 & 0.4 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.7 \\ 0.3 \end{pmatrix}$$

There is a 70% chance it will still be sunny tomorrow. You can think of this operation as the "1" in the Sunny state place flowing through the arrows of the chain one step, so that 0.7 of it remains in Sunny and 0.3 of it goes to the Cloudy state. To forecast two days out, we would do the multiplication again:

$$\begin{pmatrix} 0.7 & 0.6 \\ 0.3 & 0.4 \end{pmatrix}^2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.7 & 0.6 \\ 0.3 & 0.4 \end{pmatrix} \begin{pmatrix} 0.7 & 0.6 \\ 0.3 & 0.4 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.7 & 0.6 \\ 0.3 & 0.4 \end{pmatrix} \begin{pmatrix} 0.7 \\ 0.3 \end{pmatrix} = \begin{pmatrix} 0.67 \\ 0.33 \end{pmatrix}$$

So, if today is sunny, the odds that it will rain the day after tomorrow are 33%.

2 Modelling the Internet as a Markov Chain

We can model the web as a Markov chain, where pages that hyperlink to each other are connected by transitions. We'll assume that every person will stay on the page where they are with probability 0.5, and that the probability 0.5 that they follow a hyperlink is split evenly between every hyperlink on the page.

Let's start with a very simple model, and assume there are only four web pages on the whole internet:

- `colorado.edu`
- `colorado.edu/map`
- `colorado.edu/amath`
- `colorado.edu/amath/2360`[1]

`colorado.edu` has links to `colorado.edu/map` and `colorado.edu/amath`.
`colorado.edu/map` links to `colorado.edu`.
`colorado.edu/amath` links to `colorado.edu`, `colorado.edu/map` and `colorado.edu/amath/2360`.
`colorado.edu/amath/2360` links to `colorado.edu/amath`.

2.1 Questions

1. Sketch the Markov Chain which describes the simple internet model laid out above. You can use a tool like powerpoint, or draw it by hand, scan, and insert it into your report.
2. Suppose a user starts on the homepage, `colorado.edu`. What are the probabilities they will be on each page after one iteration?
3. Construct the transition matrix which describes the Markov Chain above. HINT: The diagonals of your matrix should be 0.5 (the odds you stay on that page) and each column should sum to one.

3 Stationary Probability Distributions

One benefit of using Markov Chains as a model for real-world phenomena is that they can provide insight into what happens as time runs to infinity. If we run the weather Markov Chain from the introduction over and over, we will find that we will eventually obtain the odds that a given day is sunny without any conditioning on what the previous day was like. This is because (under some assumptions, like that the Markov Chain is not two disjoint pieces) the system has a single stable fixed point.

[1] The true address is `https://www.colorado.edu/amath/appm-2360-intro-differential-equations-linear-algebra-fall-2018`, but that seems like a hassle to write

We think about this as multiplying the Transition matrix by itself over and over, an infinite number of times:

$$\begin{pmatrix} 0.7 & 0.6 \\ 0.3 & 0.4 \end{pmatrix}^{\infty} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \equiv \lim_{n\to\infty} \begin{pmatrix} 0.7 & 0.6 \\ 0.3 & 0.4 \end{pmatrix}^{n} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 2/3 \\ 1/3 \end{pmatrix}$$

no matter what a_1 and a_2 are, so long as $a_1, a_2 \geq 0$ and $a_1 + a_2 = 1$. This vector, $\langle 2/3, 1/3 \rangle$, is the fixed point, i.e. the stationary probability distribution. So, if you don't know anything about the past weather, your best assumption would be that it is twice as likely to be sunny as cloudy.

It turns out that if you know the structure of your transition matrix, you don't need to do expensive calculations to determine the stationary distribution, as we will see.

3.1 Questions

1. Assuming you start on the homepage `colorado.edu` with probability 1, perform 100 steps of the Markov chain. Plot the probability you are on each of the four pages at a given iteration on a single plot.
   NOTE: It may be useful to, at each step, divide your vector by the sum of its entries. If your computer was infinitely precise, this wouldn't be necessary, as each column of your matrix sums to one and it should preserve the total magnitude. However, your computer *is not* infinitely precise, and depending on rounding error of your entries your vector could erroneously grow or decay in magnitude (While the sum of all transition probabilities from a given state must be always exactly 1).

2. Make an argument that you now know the stationary distribution to within a reasonable error. What is it?

3. Recall that the set of all eigenvectors of an invertible $n \times n$ real matrix form a basis for the $\mathbb{R}^n$. That is, for eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$ and corresponding eigenvectors $v_1, v_2, ..., v_n$ any real n-dimensional vector x can be expressed as $x = c_1 v_1 + c_2 v_2 + ... + c_n v_n$. Using this fact, describe a relationship between the stationary distribution of a Markov chain and the eigenvectors or values of the corresponding transition matrix.
   Hint: the largest eigenvalue of a transition matrix is always $\lambda_1 = 1$.

4. Suppose now that a fifth page is added, `colorado.edu/project2.pdf`. Suppose `colorado.edu/amath/2360` links to this page, but the project page doesn't link to *anything*, so the probability someone who goes there will stay there is $P = 1$ instead of $P = 0.5$. What is the new transition matrix according to the rules stated in the beginning of section 2 ? Using a modified version of your code, calculate the new stationary distribution and include the corresponding plot. It may be helpful to use more iterations; keep going until the distribution converges. What happened? Describe how and why the new state changed the stationary distribution.

5. By comparing the vector in the previous part against the analytically known stationary distribution, create a `semilogy` error plot which has the log of absolute error on the y axis and iterations on the x axis. What does this tell you about the Markov chain's convergence toward the stationary distribution?
   HINT: To find the analytic stationary distribution, look at single vector in the nullspace of $A - I$.

4 Application: Page Ranking

When you make a search on an engine like Google, the search engine needs to decide what order to present results to you in. One algorithm is to assume that pages which are frequently linked to are more credible or otherwise relevant. Thus, a common component in these ranking algorithms is the determination of a stationary distribution for the Internet's Markov chain. However, one must be careful, as a blind implementation of this kind of algorithm can cause its own issues.

4.1 Questions

1. How would a stationary-distribution-based algorithm rank the importance of each of the four pages in the first part of the previous section?

2. What about the five pages in the last part? Is this a reasonable ranking ?

5 Report Guidelines

Your group will submit your project write-up on Canvas to the appropriate "Project Assignment" (you can find these under the "assignments" tab in Canvas). Adhere to the following guidelines:

- Do not put off finding a group (you must works in groups of 2-3). You should have a group set-up within one week of the project assignment due date.
- Submit your project in a **pdf** format and submit ALL code used for your project (.nb files for Mathematica, .m files for Matlab, .py or ipynb for python). Code in **Matlab, Mathematica, Maple, R, python or Julia** is acceptable (Matlab is recommended). Code in *Microsoft Word or Excel* (or any other spreadsheet program) is *not acceptable.* All other languages need instructor permission (please ask as soon as possible).

  Code may be included in the appendix if you wish. DO NOT submit anything on Canvas as a .zip file. Contents of .zip files will not necessarily be graded.
- Have only ONE group member submit the project. Having multiple people in your group submit the project to Canvas will result in multiple grades, and we will take the LOWEST one.
- Include the names and recitation section numbers of all group members working on the project on the cover page of the report.
- When you submit the report to Canvas, please include each group member's information (name, student number, and section number) in the comments. This allows us to quickly search for a student's report.

Your report needs to accurately and consistently describe the steps you took to answer the posed questions. This report should have the look and feel of a technical paper. Presentation and clarity are very important. Here are some important items to remember:

- Remember: you are to submit a complete report for this project. Documents submitted with numbered responses will be severely penalized.
- Labs must be typed, including all equations in the main body (part of your learning experience is to learn how to use an equation editor). An exception can be made for lengthy calculations in the appendix, which may be hand written (as long as they are neat and clear), and minor labels on plots, arrows in the text, and a few subscripts.
- Write your report in an organized and logical fashion. Section headers such as Introduction, Background, Problem Statement, Calculations, Results, Conclusion, . . . are not mandatory but are highly recommended. They not only help you write your report, but help the reader navigate your paper.
- Start with an introduction that describes what you will discuss in the body of the report. A brief summary of important concepts used in your discussion could be helpful here as well. Always introduce relevant equations that will be used or discussed in the report.
- Always include units in your answers
- Always label plots and refer to them in the text.
- You must include any plot that supports your conclusions or gives you insight in your investigations. However, DO NOT use screen-shots of your figures.
- DO NOT include printouts of computer software screens. You simply need to state which software you used in each step and what it did for you.

- The main body of your paper should NOT include lengthy calculations. These should be included in a labeled appendix and should be referred to in the main body.
- DO include the results of any calculations in the body of your report.
- Your report does not have to be long. You need quality, not quantity, of work. Do not omit any important piece of information, but do not feel obligated to add any extras.
- Summarize what you have accomplished in the conclusion. No new information or new results should appear in your conclusion. You should only review the highlights of what your wrote about in the body of the report.

Examples of lab reports

- Here is an example of an old good lab report which would receive high marks
- Here is an example of an old bad lab report which would receive low marks; note that in the "bad lab report", all questions are technically answered, but not in the form of a lab report