Summary of Last Lecture

A Markov Chain is a series of random variables, such that

- We know the distribution of X_0, denoted as π_0.
- We know the transition probability matrix P.
- $P\left(X_{next} = j \middle| X_{prev} = i\right) = P(i,j)$
- The distribution of kth random variable in the chain is $\pi_k = \pi_0 P^k$

By modeling a problem as a Markov Chain, can solve using "first step analysis".

- Computations are often easier than other "lower level" techniques.
- Examples:
    - Finding the expected time to reach a given state (getting two tails in a row).
    - Finding the probability of reaching one state before another (making $100 before running out of money).

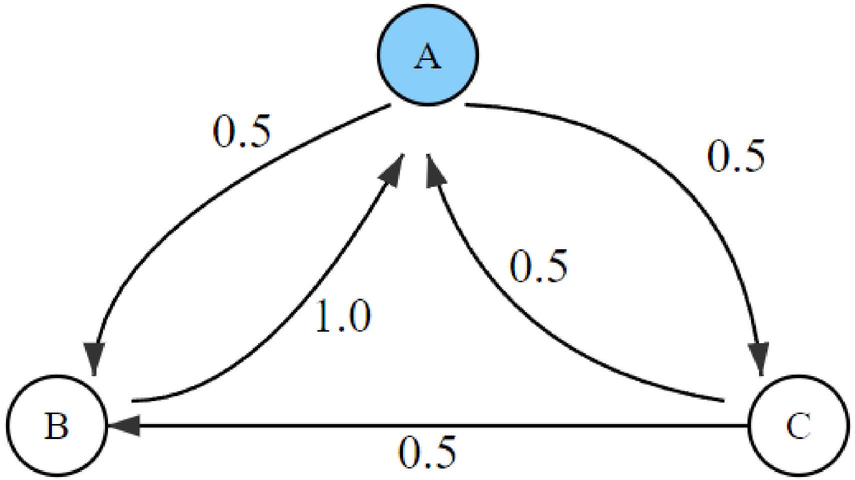
The Stationary Distribution Revisited

Lecture 27, CS70 Summer 2025

Long Term Behavior of Three State Markov Chain

Consider the Markov Chain below.

$$P = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 0.5 & 0.5 & 0 \end{bmatrix}$$



Statistics

State	Count	Fraction
A	716	0.4420
B	545	0.3364
C	359	0.2216

If we generate samples, we end up with around:

- 44% of the time in state A.
- 33% of the time in state B.
- 22% of the time in state C.

Long Term Behavior of Three State Markov Chain using Linear Algebra

We can use basic linear algebra to show why we always end up with this 44%/33%/22% pattern.

Let π_0 be a vector giving the probability that we're in any given state. For our simulation, that means $\pi_0 = [1 \ \ 0 \ \ 0]$, then $\pi_1 = [0 \ \ 0.5 \ \ 0.5]$.

- $\pi_1 = \pi_0 P = [0 \ \ 0.5 \ \ 0.5]$
- $\pi_2 = \pi_1 P = [0.75 \ \ 0.25 \ \ 0]$
- $\pi_3 = \pi_2 P = [0.25 \ \ 0.375 \ \ 0.375]$
- $\pi_4 = \pi_3 P = \pi_2 P^2 = \pi_1 P^3 = \pi_0 P^4 = [0.5625 \ \ 0.3125 \ \ 0.125]$
- $\pi_5 = \pi_0 P^5 = [0.375 \ \ 0.34375 \ \ 0.28125]$
- …
- $\pi_9 = \pi_0 P^9 = [0.4375 \quad 0.33398438 \quad 0.22851562]$

Long Term Behavior of Three State Markov Chain using Linear Algebra

We can use basic linear algebra to show why we always end up with this 44%/33%/22% pattern.

Let π_0 be a vector giving the probability that we're in any given state. For our simulation, that means $\pi_0 = [1 \;\; 0 \;\; 0]$, then $\pi_1 = [0 \;\; 0.5 \;\; 0.5]$.

- $\pi_n = [1 \;\; 0 \;\; 0]P^n$

Limit as $n \to \infty$ is $\begin{bmatrix} \frac{4}{9} & \frac{3}{9} & \frac{2}{9} \end{bmatrix}$

In this lecture, we'll show that this holds for any starting distribution, not just $\pi_0 = [1 \;\; 0 \;\; 0]$.

Example: Stationary Distribution

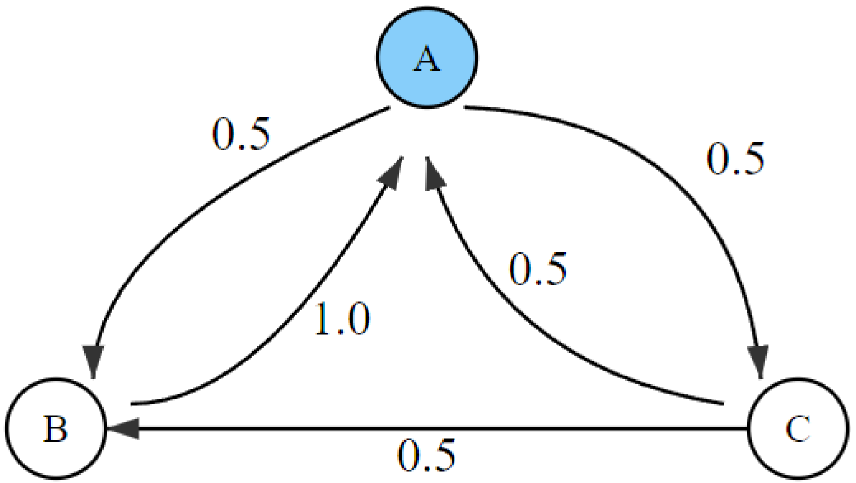
A distribution π over $\mathfrak{X} = \{1, 2, \ldots, K\}$ is stationary (a.k.a. invariant) if $\pi = \pi P$.

Note: "Stationary is forever": If $\pi_0 = \pi_1 = \pi_0 P$ *then* $\pi_0 = \pi_1 = \pi_2 = \cdots = \pi_n = \cdots$

Example: For our three state Markov Chain, $\pi = \begin{bmatrix} \frac{4}{9} & \frac{3}{9} & \frac{2}{9} \end{bmatrix}$ is stationary.

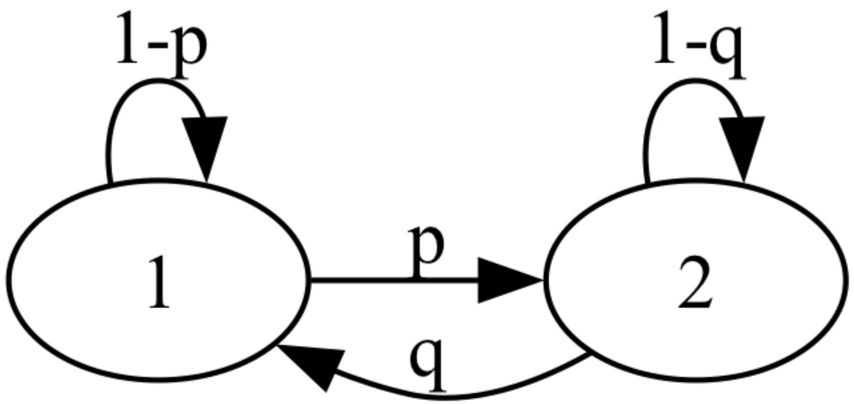
Proof: $\begin{bmatrix} \frac{4}{9} & \frac{3}{9} & \frac{2}{9} \end{bmatrix} \begin{bmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 0.5 & 0.5 & 0 \end{bmatrix}$

- $4/9 \times 0 + 3/9 \times 1 + 2/9 \times 0.5 \quad = 4/9$
- $4/9 \times 0.5 + 3/9 \times 0 + 2/9 \times 0.5 \quad = 3/9$
- $4/9 \times 0.5 + 3/9 \times 0 + 2/9 \times 0 \quad = 2/9$



Example 2: Stationary Distribution

Consider this Markov Chain:



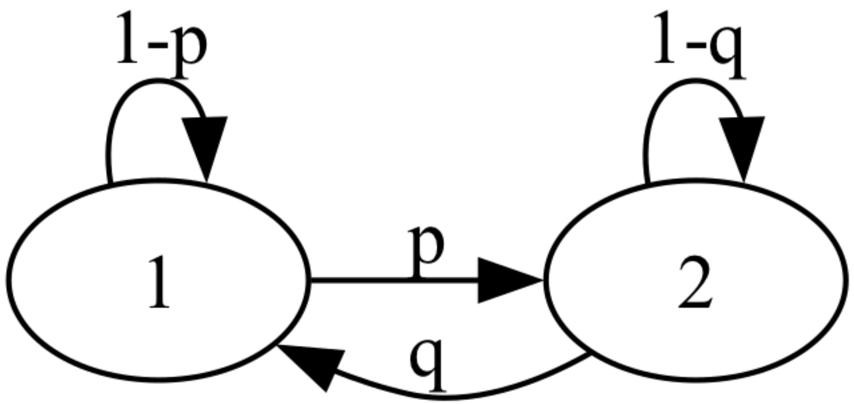
$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

π = [a b] is stationary iff $[a\ b] = [a\ b] \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$

$$a = a(1-p) + bq \iff ap = bq$$

Example 2: Stationary Distribution

Consider this Markov Chain:



$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

π = [a b] is stationary iff $[a\ b] = [a\ b] \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$

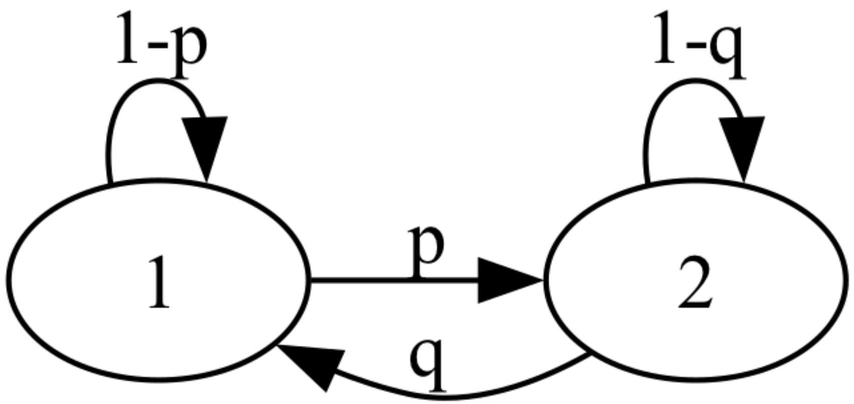
$$a = a(1-p) + bq \iff ap = bq$$

$$b = ap + b(1-q) \iff bq = ap$$

What we do next? We have two equations and two unknowns, but the equations are redundant. Are there other constraints we're not using?

Example 2: Stationary Distribution

Consider this Markov Chain:



$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

π = [a b] is stationary iff $[a\ b] = [a\ b] \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$

$$a = a(1-p) + bq \iff ap = bq$$

$$b = ap + b(1-q) \iff bq = ap$$

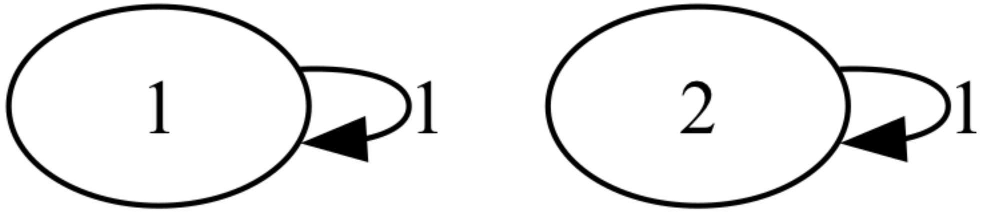
We also know: $a + b = 1$

$$\pi = \left[\frac{q}{p+q}, \frac{p}{p+q}\right]$$

Example 3: Stationary Distribution

Consider this Markov Chain:

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



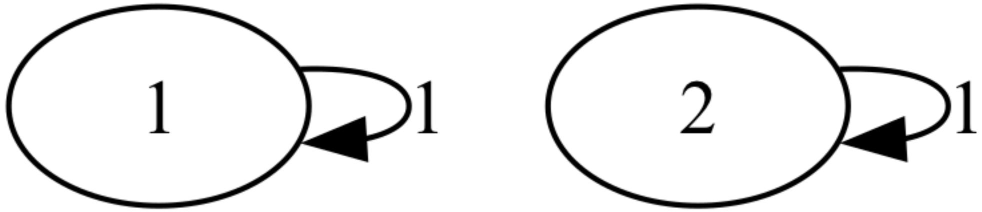
$\pi = [\mathrm{a\ b}]$ is stationary iff $[a\ b] = [a\ b]\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

What distribution is stationary, i.e., what a and b can we pick?

Example 3: Stationary Distribution

Consider this Markov Chain:

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



$\pi = [\mathrm{a\ b}]$ is stationary iff $[a\ b] = [a\ b] \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

What distribution is stationary, i.e., what a and b can we pick?

- Any distribution, e.g., $\pi = [0.3\ \ 0.7]$ just ends up right back where you started (and so does anything else).

Stationary distribution exists, but *not unique*.

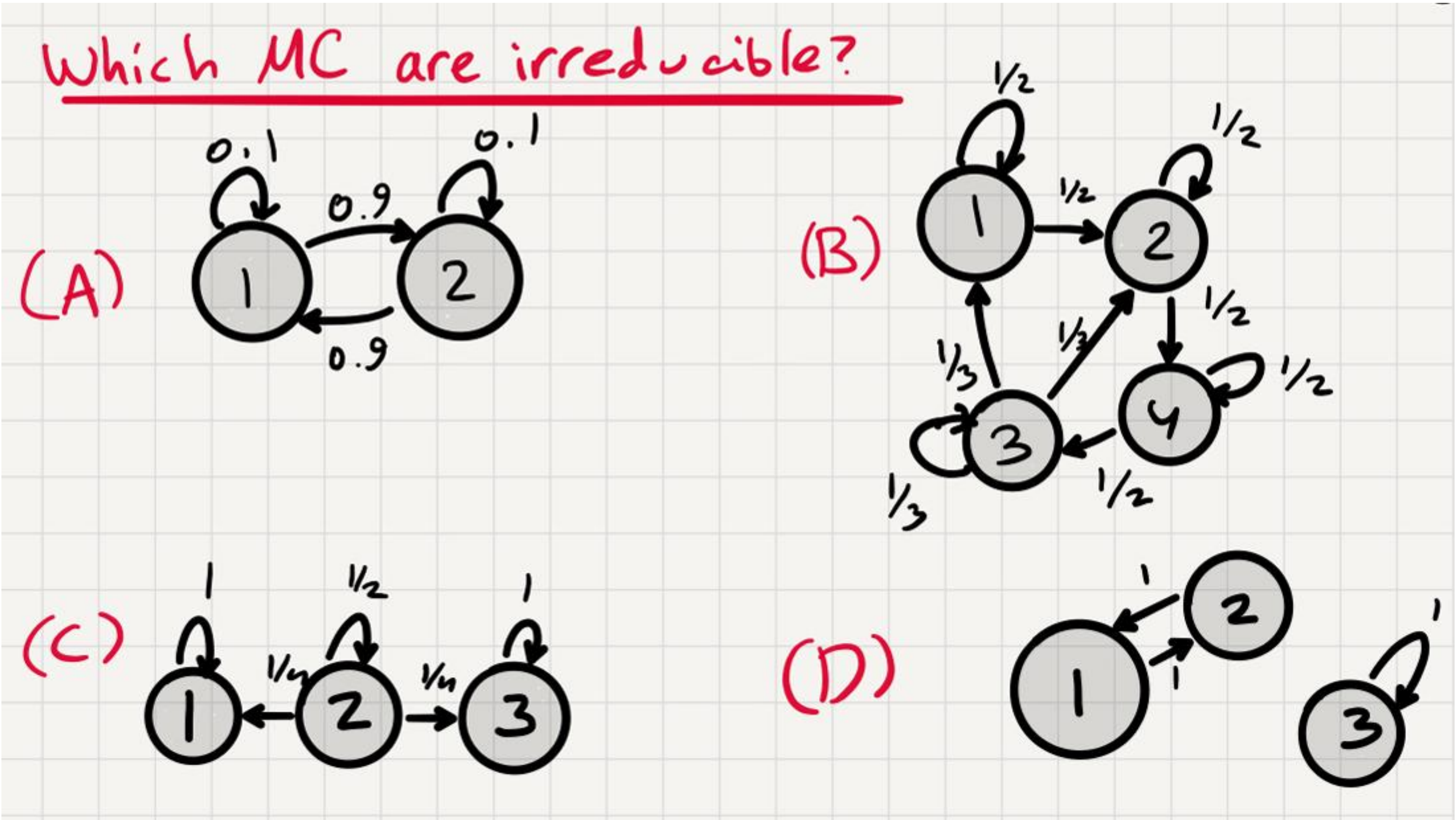
Irreducibility and Periodicity

Lecture 27, CS70 Summer 2025

Irreducible Markov Chains

A Markov Chain is irreducible if you can get from any state i to any other state j, possibly by following multiple (non-zero probability) transitions.

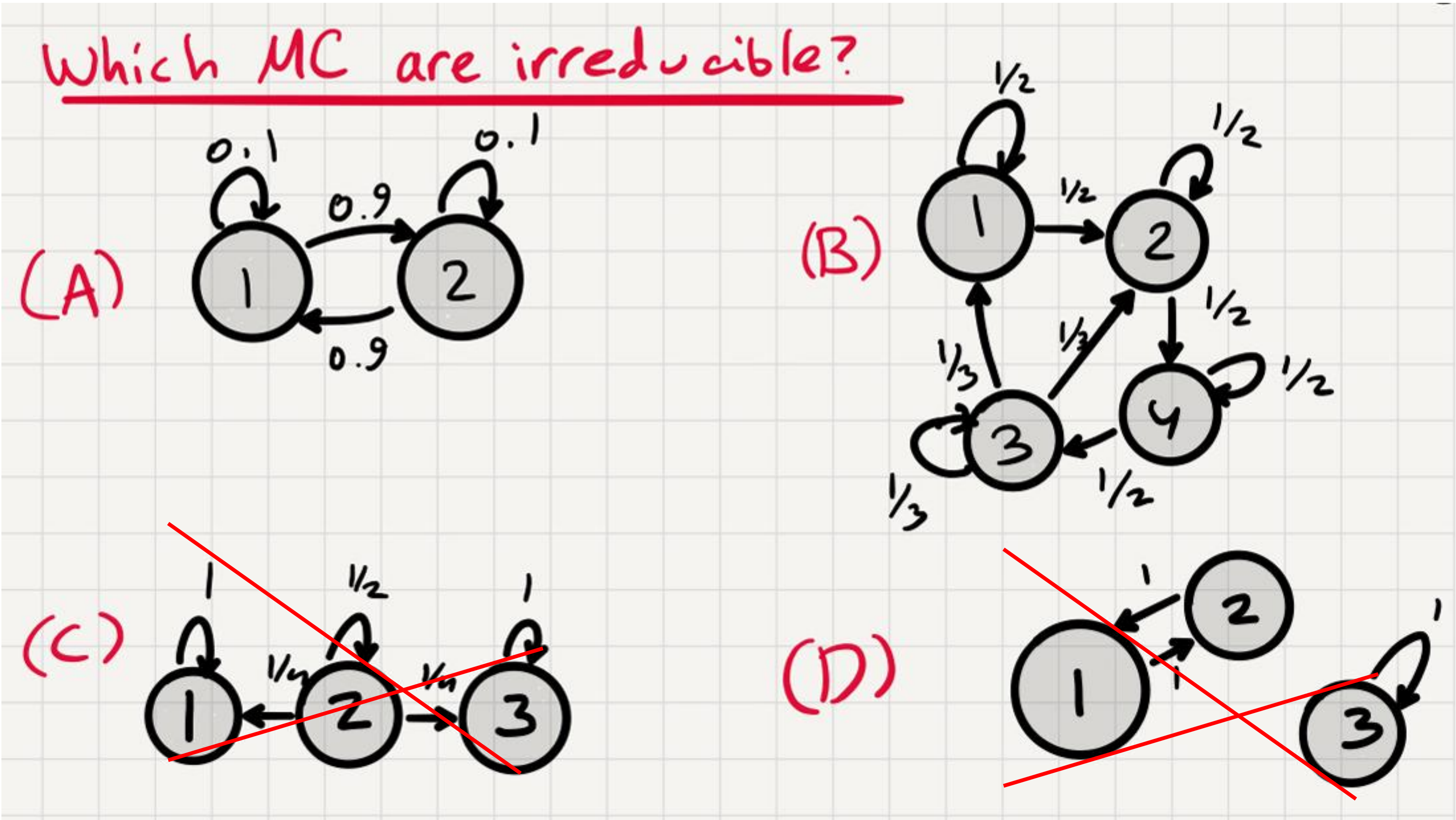
Stealing a nice exercise from Avishay Tal in Fall 2023:



Irreducible Markov Chains

A Markov Chain is irreducible if you can get from any state i to any other state j, possibly by following multiple steps.

Stealing a nice exercise from Avishay Tal in Fall 2023:



A few side notes about irreducibility

A Markov Chain is irreducible if you can get from any state i to any other state j, possibly by following multiple steps.

Visually checking: Can you find a cycle through all vertices?

- Not restricting number of times using a vertex (not necessarily Hamiltonian)

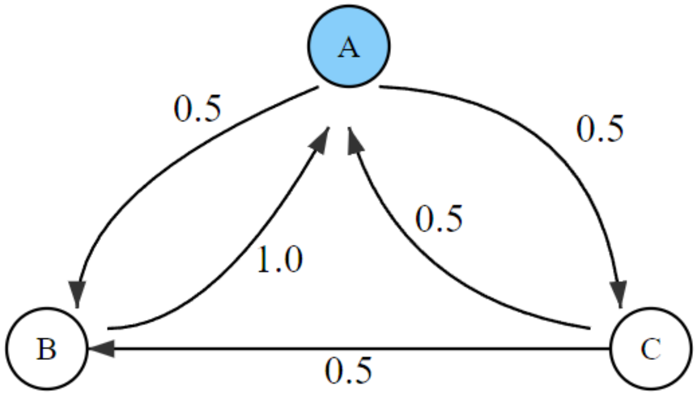
Graph theory connection:

- Treating as a directed graph including only edges with >0 probability...
- Such a graph is called a strongly connected graph (path from any i to any j)
- There are linear time algorithms (fast!) to test if a graph is strongly connected

Unique stationary distribution exists iff Markov Chain is irreducible

Theorem: A Markov Chain has a unique stationary distribution π with $\pi(i) > 0$ for all states i if and only if the Markov Chain is irreducible.

Our ongoing example:



Irreducible? Yes!

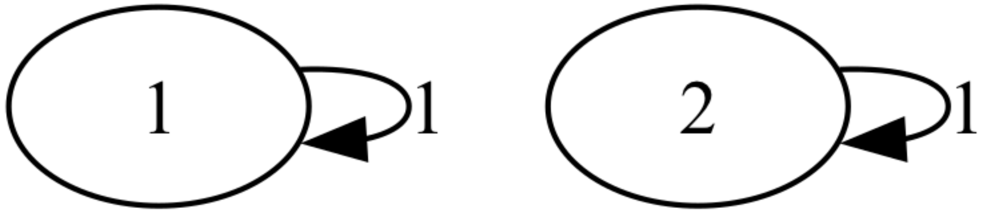
Stationary distribution? Yes! Recall from earlier: $\pi = \begin{bmatrix} \frac{4}{9} & \frac{3}{9} & \frac{2}{9} \end{bmatrix}$

Theorem tells us: This is the only stationary distribution.

Unique Stationary Distribution – Example 2

Theorem: A Markov Chain has a unique stationary distribution π with $\pi(i) > 0$ for all states i if and only if the Markov Chain is irreducible.

Another previous example:



Irreducible? No!

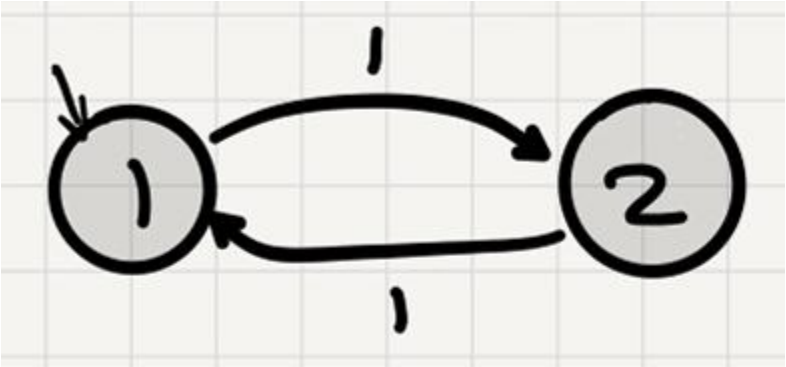
Stationary distribution? Yes! $\pi = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix}$ But also $\pi = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \end{bmatrix}$ and others....

Theorem is consistent with these examples. Won't prove it, but this is reassuring.

Unique Stationary Distribution – Example 3

Theorem: A Markov Chain has a unique stationary distribution π with $\pi(i) > 0$ for all states i if and only if the Markov Chain is irreducible.

A new example:



Irreducible? Yes!

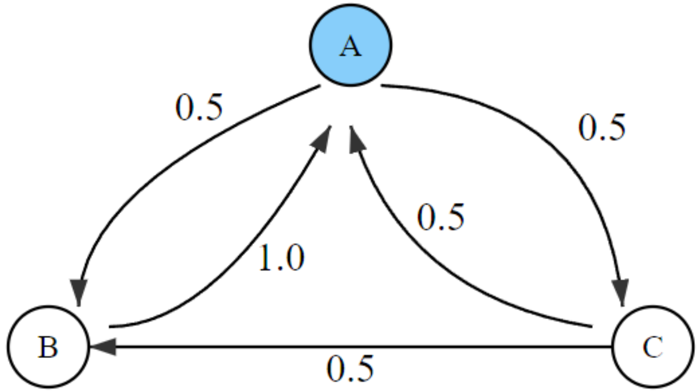
Stationary distribution? Yes! $\pi = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix}$

Theorem tells us: This is the only stationary distribution. But something different...

Convergence?

Recall from earlier:

Let π_0 be a vector giving the probability that we're in any given state. For our simulation, that means $\pi_0 = [1 \;\; 0 \;\; 0]$, then $\pi_1 = [0 \;\; 0.5 \;\; 0.5]$.



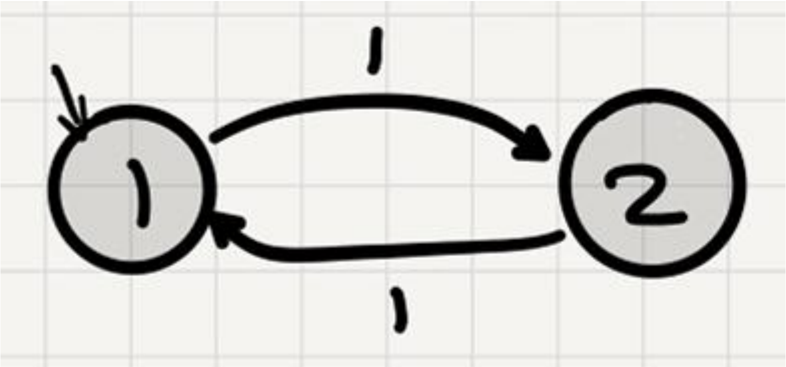
- $\pi_1 = \pi_0 P = [0 \;\; 0.5 \;\; 0.5]$
- $\pi_2 = \pi_1 P = [0.75 \;\; 0.25 \;\; 0]$
- $\pi_3 = \pi_2 P = [0.25 \;\; 0.375 \;\; 0.375]$
- $\pi_4 = \pi_3 P = \pi_2 P^2 = \pi_1 P^3 = \pi_0 P^4 = [0.5625 \;\; 0.3125 \;\; 0.125]$
- $\pi_5 = \pi_0 P^5 = [0.375 \;\; 0.34375 \;\; 0.28125]$
- …
- $\pi_9 = \pi_0 P^9 = [0.4375 \quad 0.33398438 \quad 0.22851562]$

Fact: Distribution converges to stationary dist. Proof: "Trust me"

Convergence?

What about other Markov Chain?

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



Now with $\pi_0 = [1\ 0]$:

- $\pi_1 = \pi_0 P = [1\ 0] \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = [0\ 1]$
- $\pi_2 = \pi_1 P = [0\ 1] \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = [1\ 0]$
- ... and we're back where we started!
- $\pi_3 = [0\ 1]$
- $\pi_4 = [1\ 0]$
- $\pi_5 = [0\ 1]$

.....

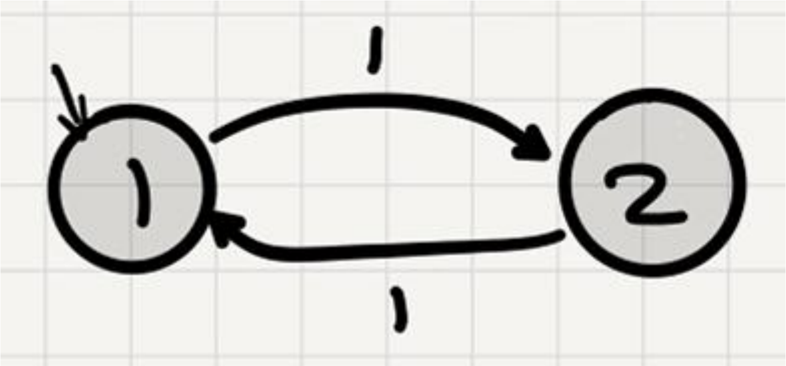
So:

Unique stationary distribution (by theorem)

But doesn't converge to it for some choices of π_0

Periodicity

The periodicity of a Markov Chain with transition matrix P is the gcd of the lengths of all closed walks (cycles) in the chain.

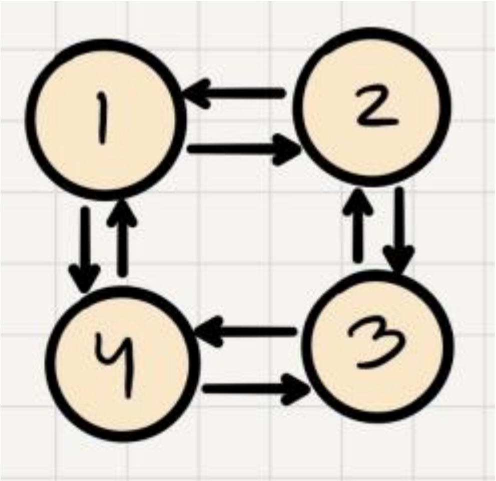
Example: 

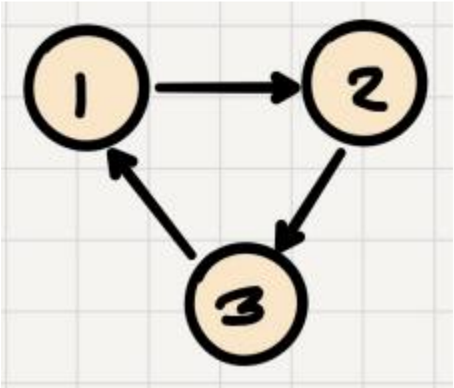
gcd of lengths of all cycles, which is 2.

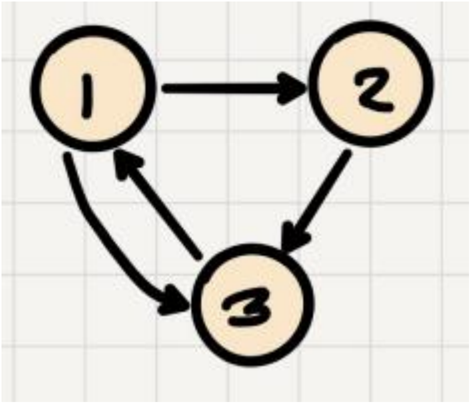
Periodicity

The periodicity of a Markov Chain with transition matrix P is the gcd of the lengths of all closed walks (cycles) in the chain.

More examples: What do you think are the periods of these graphs?



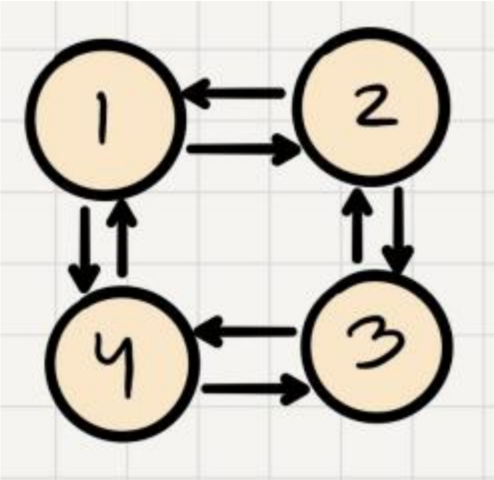




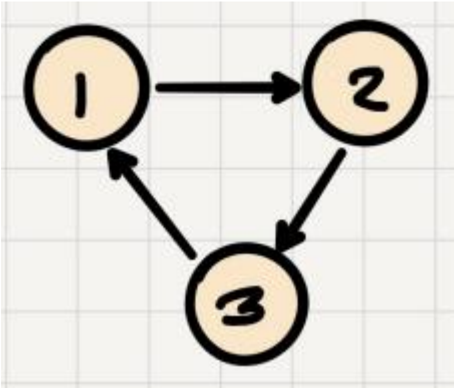
Periodicity

The periodicity of a Markov Chain with transition matrix P is the gcd of the lengths of all closed walks (cycles) in the chain.

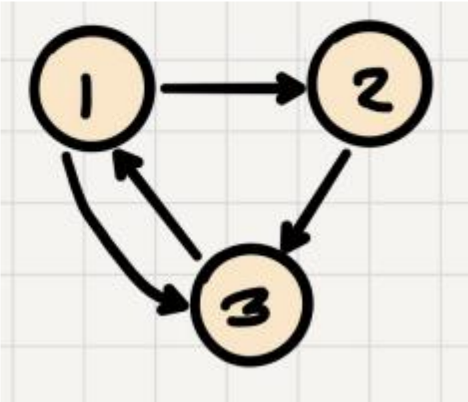
More examples: What do you think are the periods of these graphs?



Period: gcd(2, 4)=2



Period: 3



Period: gcd(2, 3)=1

Period = 1 is the important case! Such a Markov chain is aperiodic.

Why Period Is Important

Theorem: Let $c_1, c_2, \dots, c_k$ be cycle lengths in a directed graph such that $\gcd(c_1, c_2, \dots, c_k) = 1$. Then there is an n_0 such that for any vertex v and all $n \geq n_0$ there is a tour of length n from v back to itself.

Related to "Making Change" problem (coin denominations, ...)

What it means here: Given these cycle lengths we can get back to a vertex in <u>any number of steps</u> as long as n is big enough (and we're interested in large n).

Conversely: If $\gcd(c_1, c_2, \dots, c_k) = 2$ (for example), every path back will have even length – *no path back of odd length is possible!*

$\Longrightarrow$ *If we start in state s, $\pi_i(s)$ will be non-zero if i is even, zero if i is odd… No convergence*

Tip: If the Markov Chain has any self-loops (which is common), it is aperiodic!

The Fundamental Theorem of Markov Chains

We've seen some Markov Chains that converge to some unique invariant distribution, and others that do not.

The Fundamental Theorem of Markov Chains: If a Markov Chain is finite, irreducible, and aperiodic, then for any initial π_0 the distribution at time n converges as $n \to \infty$ to π, which is the unique invariant distribution, and $\pi(i) > 0$ for all states i.

Irreducible: Any node reachable from any other node.

Aperiodic: Period is 1.

- Note: Must contain loops (irreducible), but GCD of the length of those loops is 1.

An Application: Predicting Text

Lecture 27, CS70 Summer 2025

In 1948, Claude Shannon wrote a paper called "A Mathematical Theory of Communication".

- Focus: What is "information" and how do we encode it?

Why? (other than it's an incredibly fundamental question)

Data compression in half a slide:

Predict characters:	Use to encode:
P("A")=1/8	A = 111
P("B")=1/2	B = 0
P("C")=1/8	C = 110
P("D")=1/4	D = 10

$$3 \cdot 1/8 + 1 \cdot 1/2 + 3 \cdot 1/8 + 2 \cdot 1/4 = 1.75$$

Claude Shannon: A Mathematical Theory of Communication (1948)

In 1948, Claude Shannon wrote a paper called "A Mathematical Theory of Communication".

- Focus: What is "information" and how do we encode it?

How to predict?

Order 0: A single probability distribution for each letter

E = 12.7%
T=9.1%
A=8.2%
O=7.5%
...

Remarkably consistent across a wide variety of English writing...

Sufficient to encode English text in about 4.2 bits/letter

Claude Shannon: A Mathematical Theory of Communication (1948)

In 1948, Claude Shannon wrote a paper called "A Mathematical Theory of Communication".

- Focus: What is "information" and how do we encode it?

How to predict?

Order 1: Probability of next letter depends on previous

$$P(C_i = \text{"H"} \mid C_{i-1} = \text{"T"}) \neq P(C_i = \text{"H"} \mid C_{i-1} = \text{"Q"})$$

Context matters… predictions are more accurate considering previous letter

This is a Markov Chain, where the state is the last letter seen!

Claude Shannon: A Mathematical Theory of Communication (1948)

In 1948, Claude Shannon wrote a paper called "A Mathematical Theory of Communication".

- Focus: What is "information" and how do we encode it?

How to predict?

Order 2: Probability of next letter depends on *two* previous ("digrams")

$$P(D_i = \text{"ON"} \mid D_{i-1} = \text{"IO"}) \neq P(D_i = \text{"ON"} \mid D_{i-1} = \text{"FO"})$$

Context matters... predictions are more even accurate with 2 previous letters

This is a Markov Chain, where the state is the last digram seen!

Can encode at roughly 2.77 bits/letter.

Modern Compression – and more

There's an entire class of compression algorithms based on Markov Chain predictions: the "PPM" (Prediction by Partial Matching) algorithms.

In the real world: The RAR compression program uses this.

If you can predict, what else can you do?

Generate text!

Auto-complete

Large language models (LLMs)

Some examples of text generation

Order 0 example: Follows letter frequencies, but that's about it:

```
OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL
```

Going up to order 4 (trained on all ASCII chars, using Moby Dick):

```
I am in that down broad glazier.” true cylindefinitely from yondering again
one visible tinkled in this colourite myself to suit of the old wrinkling
that stood city offin.”
```

Better... but still not going to fool anyone.

LLMs: Token based prediction (not characters) with additional context

... but it's still just a (very large) Markov Chain!

Reflections on CS70

Lecture 27, CS70 Summer 2025

Day 1 Slide: "What Is This Class?"

Discrete Math: Math with structures with distinct objects

- Not continuous
- Not "discreet"!
- But not (necessarily) finite
- Digital? What computers work with...

Probability Theory: Probability and properties of random events

- Can use continuous functions
- Basically counting....

But really this class is about: building important ideas by putting together simple concepts; careful and precise reasoning about those constructions; proofs; counting

This can be uncomfortable at first – hopefully you feel better about it now...

Also From Day 1

Lots of people to thank:

And thank *you*!

If Berkeley students, I hope this supports your future studies!

If visitors, I hope you found your time here worthwhile

(*and that you had some time to enjoy being here*!)

You're not done yet – the final exam!

Important details:

Main exam location: Dwinelle 155

Main exam day/time: Tuesday, Aug 12, 7:00-10:00

Those with DSP accommodations: You'll be contacted with details

What topics are covered?

Everything in the notes except "optional" parts – and not my side-tracks...

Roughly 2/3 on post-midterm (counting and later) material

Taking the exam:

You can bring hand-written notes: two sheets of paper (front and back)

Nothing else!

Remember to write answers *inside the marked boxes*.

Final exam - Preparation

Best resources

- The notes (always!) and lectures
- Discussion sheets and posted HW solutions
- Past exams (do a realistic run-through!)

Final review by TAs:

- When? Fri Aug 8, 12:00-2:00 and 3:00-5:00 – 30-minute blocks for topics
- Where? Cory 521
- See Ed post for more details

Also: Regular TA office hours on Friday

Monday:

- Student-choice lecture (see Ed)
- No discussion or TA office hours – professor OH if needed

And Ed is always available for questions/discussion

Related classes at Berkeley for those who want more!

- CS170: Efficient Algorithms and Intractable Problems a.k.a. Introduction to CS Theory: Graphs, Dynamic Programming, Complexity.
- EE126: Probability in EECS: An Application-Driven Course: PageRank, Digital Links, Tracking, Speech Recognition, Planning, etc. Hands on labs with python experiments (GPS, Shazam, ...).
- CS188: Artificial Intelligence: Hidden Markov Chains, Bayes Networks, Neural Networks.
- CS189: Introduction to Machine Learning: Regression, Neural Networks, Learning, etc. Programming experiments with real-world applications.
- EE121: Digital Communication: Coding for communication and storage.
- EE223: Stochastic Control.
- EE229A: Information Theory; EE229B: Coding Theory

The End

... and they all lived happily ever after.