COMPSCI 240: Reasoning Under Uncertainty

Nic Herndon and Andrew Lan

University of Massachusetts at Amherst

Spring 2019, Section 01

Lecture 33: Review for Final Exam

Topics

- Basic counting problems
- Probability
- Discrete random variables
- Midterm Exam #1
- Continuous random variables
- Central limit theorem
- Probabilistic reasoning
- Game theory
- Midterm Exam #2
- Markov chains
- Bayesian network
- Final Exam

Part I Overview

- Basic counting problems
  - Set theory: size of, subset, disjoint sets, partitions, power set, universal set, operations (complement, union, intersection)
  - Counting: permutations, k-permutations, combinations, partitions
- Probability
  - Probability axioms
  - Conditional probability (sequential model)
  - Multiplication rule
  - Total probability theorem
  - Bayes' rule
  - Independence
  - Conditional independence
- Discrete random variables
  - Probability mass function (PMF)
  - Common discrete RVs: uniform, Bernoulli, binomial, geometric, Poisson
  - Expectation and Variance + their properties (e.g., functions of RVs)
  - Multiple RVs (joint, marginal, conditional PMF; functions of two RVs, expectation and variance)

Part II Overview

- Continuous random variables
  - Probability *density* function (PDF), cumulative density function (CDF), and probability mass
  - Expectation and Variance + their properties
  - Common continuous RVs: uniform, exponential, (standard) normal/Gaussian
  - Multiple RVs (joint, marginal, conditional PDFs), covariance, correlation
- Limit theorems
  - Markov bound
  - Chebyshev bound
  - Weak law of large numbers, and convergence in probability
  - Strong law of large numbers
  - Central limit theorem
- Game theory
  - Strategies: pure, IESDS, mixed
  - Nash equilibrium
  - Zero-sum games

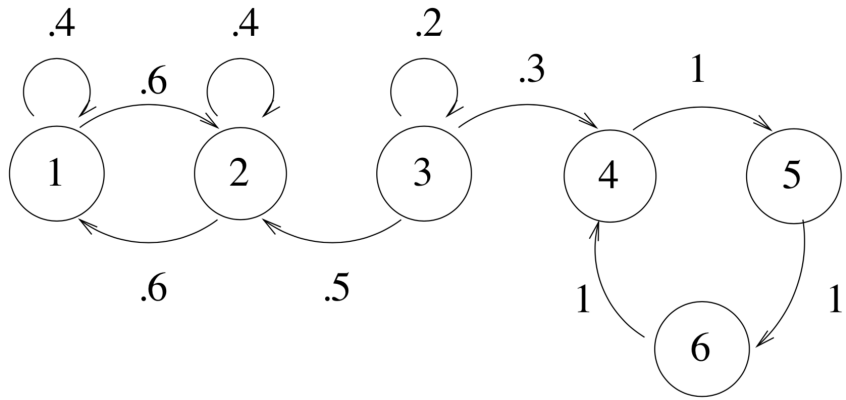
Part III Overview

- Markov chains
  - Used for problems in which the future depends on past ONLY through present
  - Consist of: state space, transition probabilities, and an initial state
  - Markov property: transition probability p_{ij} must be **non-negative** and **sum to 1**
  - Transition probability matrix
  - Markov chain theorem: $v_t = v_0 A^t$
  - Steady state distribution
  - Classification of states: recurrent, and transient; their implication on unique steady state
  - Periodic recurrent class, and steady-state convergence theorem
- Bayesian networks
  - Use known conditional independencies to factorize joint distributions using the chain rule
  - Use DAG to keep track of all conditional independence assumptions
  - The factor associated with variable X_i is $P(X_i|Pa_i)$
  - Three cases of conditional independence common in Bayesian networks
  - Types of queries: marginal, conditional, and joint
  - Estimating Bayesian networks from data

Problems from MIT OCW:
Probabilistic Systems Analysis and Applied Probability
Tutorial 9 Problems

Problem 1 (Tutorial 9 Problem 2)

The Markov chain shown below is in state 3 immediately before the first trial.



(a) Indicate which states, if any, are recurrent, transient, and periodic.

(b) Find the probability that the process is in state 3 after n trials.

(c) Find the expected number of trials up to and including the trial on which the process leaves state 3.

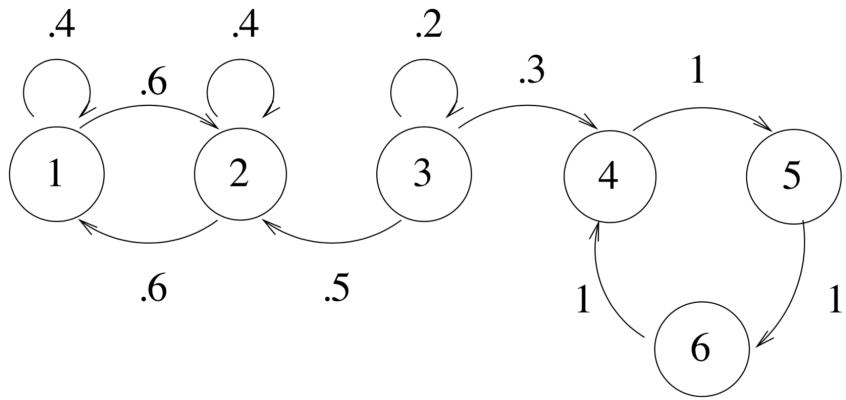
(d) Find the probability that the process never enters state 1.

(e) Find the probability that the process is in state 4 after 10 trials.

(f) Given that the process is in state 4 after 10 trials, find the probability that the process was in state 4 after the first trial.

Problem 1 (Tutorial 9 Problem 2)

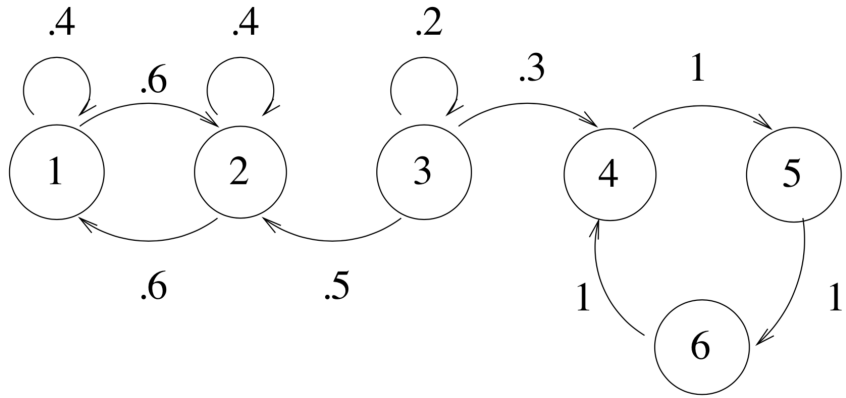
The Markov chain shown below is in state 3 immediately before the first trial.



(a) Indicate which states, if any, are recurrent, transient, and periodic.

Problem 1 (Tutorial 9 Problem 2)

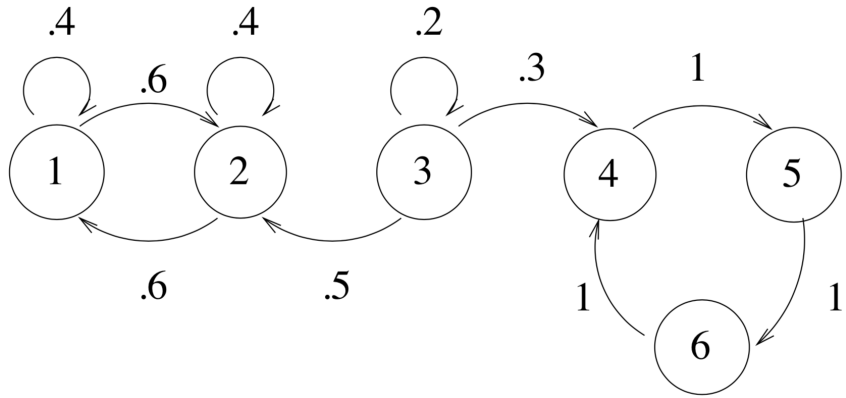
The Markov chain shown below is in state 3 immediately before the first trial.



(a) Indicate which states, if any, are recurrent, transient, and periodic.
Recurrent: 1, 2, 4, 5, 6; Transient: 3; Periodic: 4, 5, 6.

Problem 1 (Tutorial 9 Problem 2)

The Markov chain shown below is in state 3 immediately before the first trial.

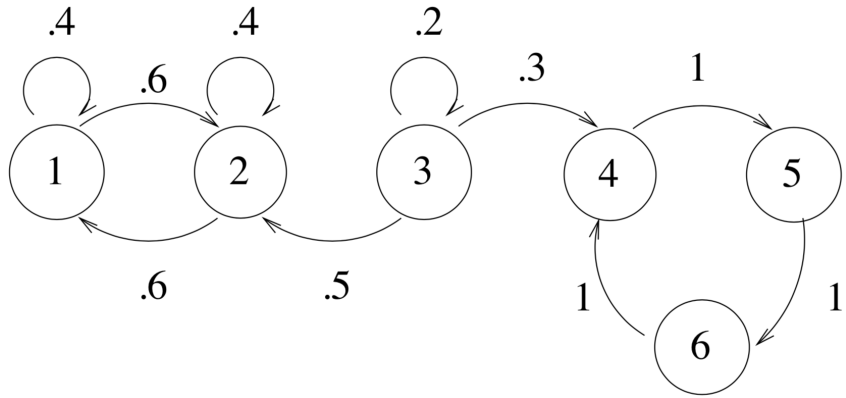


(a) Indicate which states, if any, are recurrent, transient, and periodic.
Recurrent: 1, 2, 4, 5, 6; Transient: 3; Periodic: 4, 5, 6.

(b) Find the probability that the process is in state 3 after n trials:

Problem 1 (Tutorial 9 Problem 2)

The Markov chain shown below is in state 3 immediately before the first trial.

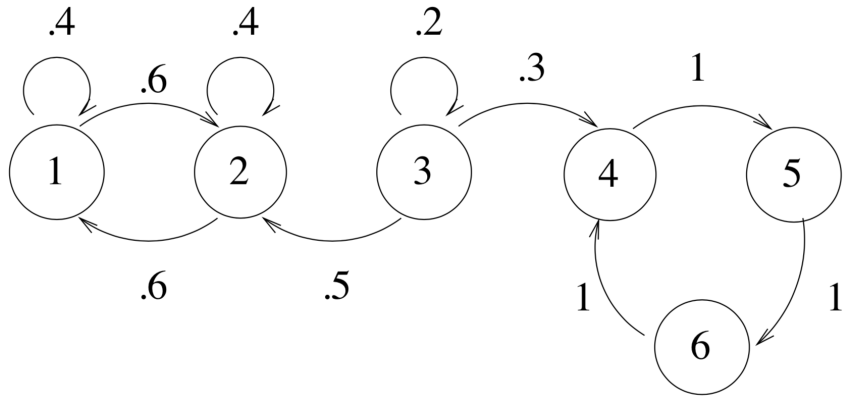


(a) Indicate which states, if any, are recurrent, transient, and periodic.
Recurrent: 1, 2, 4, 5, 6; Transient: 3; Periodic: 4, 5, 6.

(b) Find the probability that the process is in state 3 after n trials: 0.2^n

Problem 1 (Tutorial 9 Problem 2)

The Markov chain shown below is in state 3 immediately before the first trial.



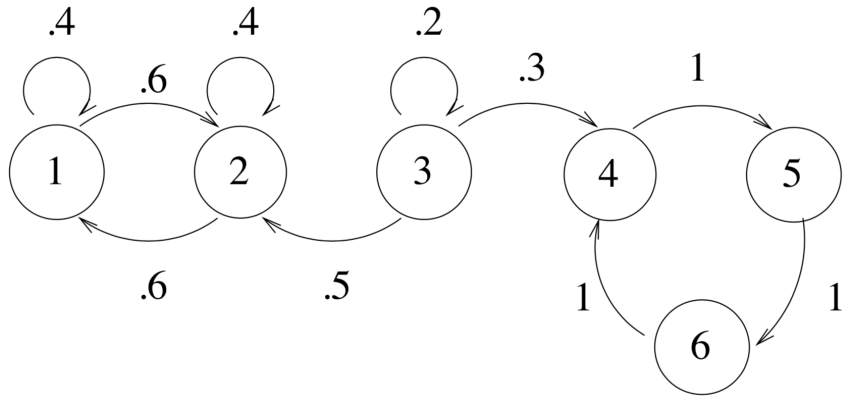
(a) Indicate which states, if any, are recurrent, transient, and periodic.
Recurrent: 1, 2, 4, 5, 6; Transient: 3; Periodic: 4, 5, 6.

(b) Find the probability that the process is in state 3 after n trials: 0.2^n

(c) Find the expected number of trials up to and including the trial on which the process leaves state 3.

Problem 1 (Tutorial 9 Problem 2)

The Markov chain shown below is in state 3 immediately before the first trial.



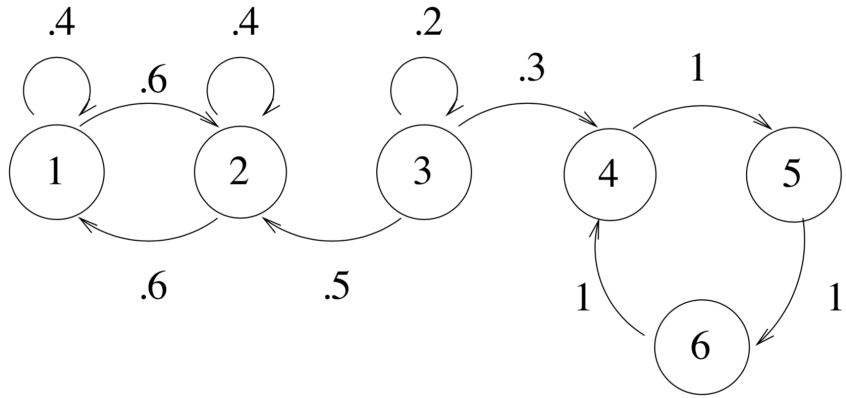
(a) Indicate which states, if any, are recurrent, transient, and periodic.
Recurrent: 1, 2, 4, 5, 6; Transient: 3; Periodic: 4, 5, 6.

(b) Find the probability that the process is in state 3 after n trials: 0.2^n

(c) Find the expected number of trials up to and including the trial on which the process leaves state 3.
This is a geometric random variable with parameter $p = 0.5 + 0.3$. Hence, the expected number of trials up to and including the trial on which the process leaves state 3 is $E[X] = 1/p = 5/4$.

Problem 1 (Tutorial 9 Problem 2)

The Markov chain shown below is in state 3 immediately before the first trial.



(a) Indicate which states, if any, are recurrent, transient, and periodic.
Recurrent: 1, 2, 4, 5, 6; Transient: 3; Periodic: 4, 5, 6.

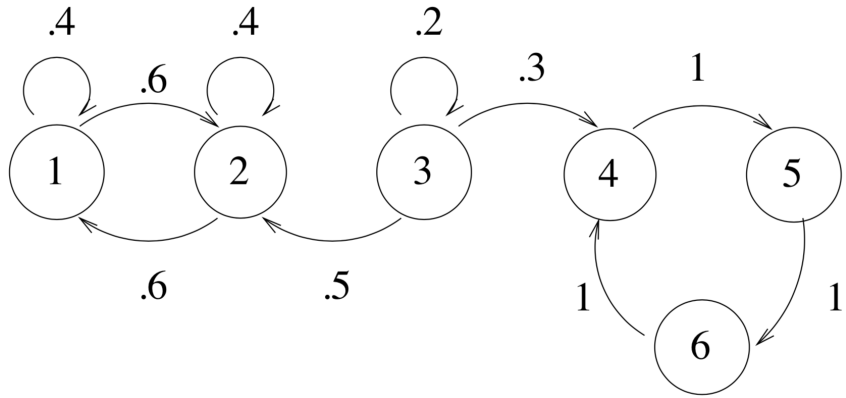
(b) Find the probability that the process is in state 3 after n trials: 0.2^n

(c) Find the expected number of trials up to and including the trial on which the process leaves state 3.
This is a geometric random variable with parameter $p = 0.5 + 0.3$. Hence, the expected number of trials up to and including the trial on which the process leaves state 3 is $E[X] = 1/p = 5/4$.

(d) Find the probability that the process never enters state 1:

Problem 1 (Tutorial 9 Problem 2)

The Markov chain shown below is in state 3 immediately before the first trial.



(a) Indicate which states, if any, are recurrent, transient, and periodic.
Recurrent: 1, 2, 4, 5, 6; Transient: 3; Periodic: 4, 5, 6.

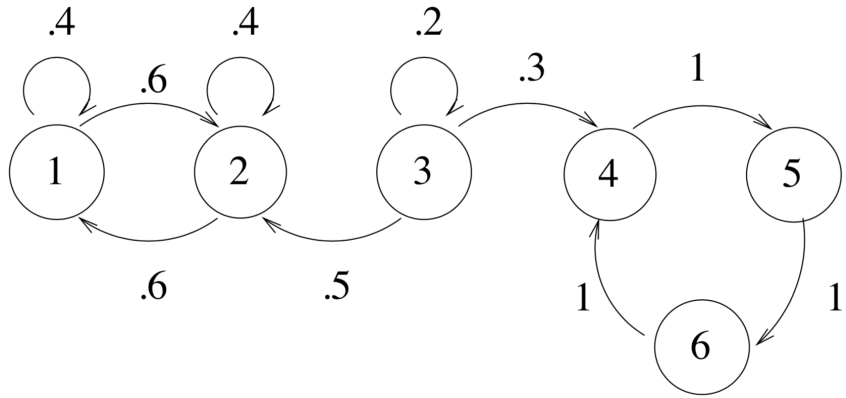
(b) Find the probability that the process is in state 3 after n trials: 0.2^n

(c) Find the expected number of trials up to and including the trial on which the process leaves state 3.
This is a geometric random variable with parameter $p = 0.5 + 0.3$. Hence, the expected number of trials up to and including the trial on which the process leaves state 3 is $E[X] = 1/p = 5/4$.

(d) Find the probability that the process never enters state 1: 3/8

Problem 1 (Tutorial 9 Problem 2)

The Markov chain shown below is in state 3 immediately before the first trial.



(a) Indicate which states, if any, are recurrent, transient, and periodic.
Recurrent: 1, 2, 4, 5, 6; Transient: 3; Periodic: 4, 5, 6.

(b) Find the probability that the process is in state 3 after n trials: 0.2^n

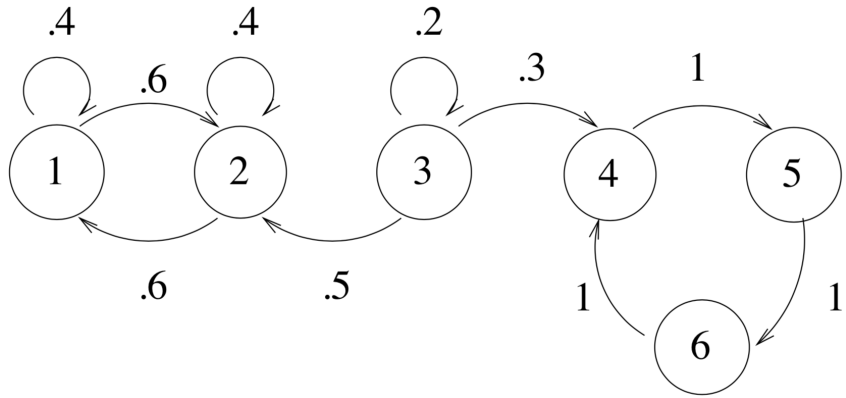
(c) Find the expected number of trials up to and including the trial on which the process leaves state 3.
This is a geometric random variable with parameter $p = 0.5 + 0.3$. Hence, the expected number of trials up to and including the trial on which the process leaves state 3 is $E[X] = 1/p = 5/4$.

(d) Find the probability that the process never enters state 1: $3/8$

(e) Find the probability that the process is in state 4 after 10 trials.

Problem 1 (Tutorial 9 Problem 2)

The Markov chain shown below is in state 3 immediately before the first trial.



(a) Indicate which states, if any, are recurrent, transient, and periodic.
Recurrent: 1, 2, 4, 5, 6; Transient: 3; Periodic: 4, 5, 6.

(b) Find the probability that the process is in state 3 after n trials: 0.2^n

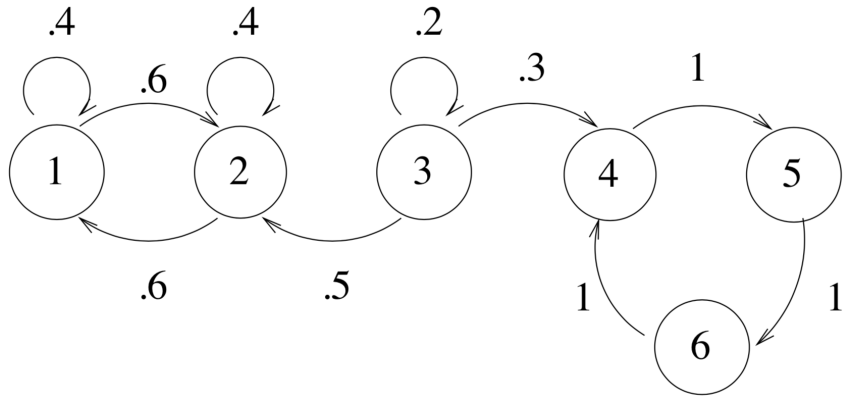
(c) Find the expected number of trials up to and including the trial on which the process leaves state 3.
This is a geometric random variable with parameter $p = 0.5 + 0.3$. Hence, the expected number of trials up to and including the trial on which the process leaves state 3 is $E[X] = 1/p = 5/4$.

(d) Find the probability that the process never enters state 1: $3/8$

(e) Find the probability that the process is in state 4 after 10 trials.
$P(A) = 0.3 + 0.2^3 0.3 + 0.2^6 0.3 + 0.2^9 0.3 = 0.3024$

Problem 1 (Tutorial 9 Problem 2)

The Markov chain shown below is in state 3 immediately before the first trial.



(a) Indicate which states, if any, are recurrent, transient, and periodic.
Recurrent: 1, 2, 4, 5, 6; Transient: 3; Periodic: 4, 5, 6.

(b) Find the probability that the process is in state 3 after n trials: 0.2^n

(c) Find the expected number of trials up to and including the trial on which the process leaves state 3.
This is a geometric random variable with parameter $p = 0.5 + 0.3$. Hence, the expected number of trials up to and including the trial on which the process leaves state 3 is $E[X] = 1/p = 5/4$.

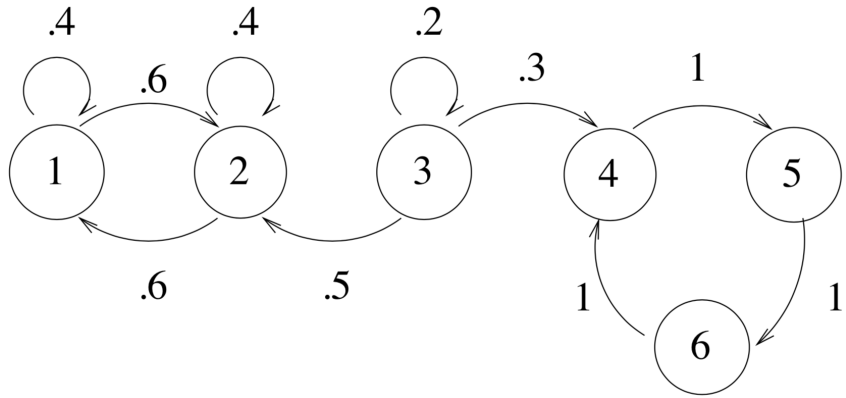
(d) Find the probability that the process never enters state 1: $3/8$

(e) Find the probability that the process is in state 4 after 10 trials.
$P(A) = 0.3 + 0.2^3 0.3 + 0.2^6 0.3 + 0.2^9 0.3 = 0.3024$

(f) Given that the process is in state 4 after 10 trials, find the probability that the process was in state 4 after the first trial.

Problem 1 (Tutorial 9 Problem 2)

The Markov chain shown below is in state 3 immediately before the first trial.



(a) Indicate which states, if any, are recurrent, transient, and periodic.
Recurrent: 1, 2, 4, 5, 6; Transient: 3; Periodic: 4, 5, 6.

(b) Find the probability that the process is in state 3 after n trials: 0.2^n

(c) Find the expected number of trials up to and including the trial on which the process leaves state 3.
This is a geometric random variable with parameter $p = 0.5 + 0.3$. Hence, the expected number of trials up to and including the trial on which the process leaves state 3 is $E[X] = 1/p = 5/4$.

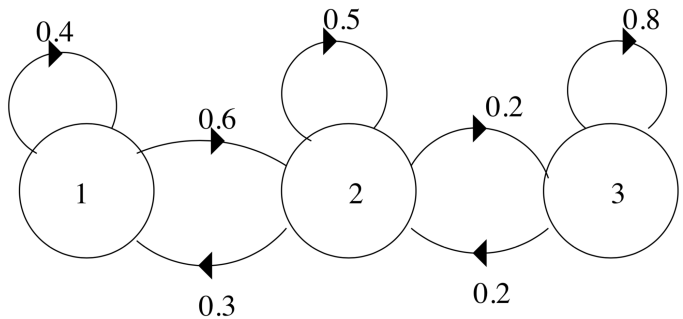
(d) Find the probability that the process never enters state 1: $3/8$

(e) Find the probability that the process is in state 4 after 10 trials.
$P(A) = 0.3 + 0.2^3 0.3 + 0.2^6 0.3 + 0.2^9 0.3 = 0.3024$

(f) Given that the process is in state 4 after 10 trials, find the probability that the process was in state 4 after the first trial.
$0.3/P(A) = 0.992$

Problem 2 (Tutorial 9 Problem 3)

Consider the Markov chain below. Let us refer to a transition that results in a state with a higher (respectively, lower) index as a birth (respectively, death). Calculate the following quantities, assuming that when we start observing the chain, it is already in steady-state.



(a) For each state i, the probability that the current state is i.

(b) The probability that the first transition we observe is a birth.

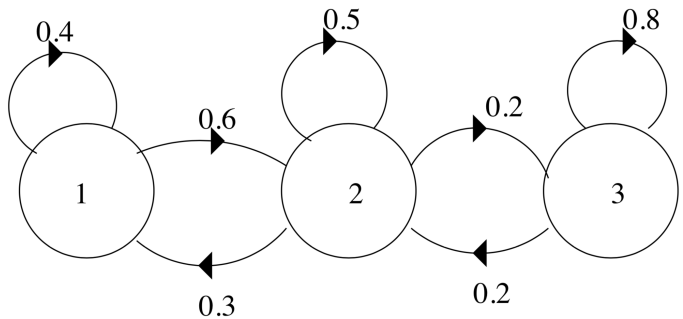
(c) The probability that the first change of state we observe is a birth.

(d) The conditional probability that the process was in state 2 before the first transition that we observe, given that this transition was a birth.

(e) The conditional probability that the process was in state 2 before the first change of state that we observe, given that this change of state was a birth.

Problem 2 (Tutorial 9 Problem 3)

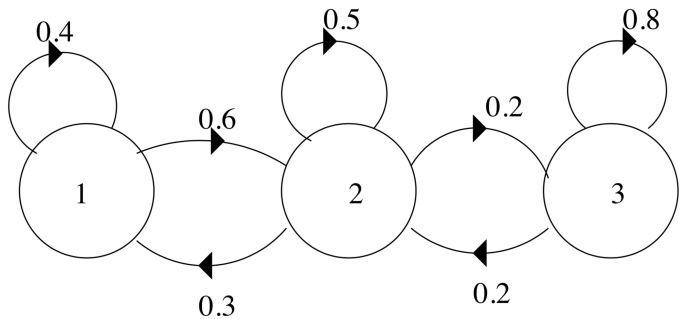
Consider the Markov chain below. Let us refer to a transition that results in a state with a higher (respectively, lower) index as a birth (respectively, death). Calculate the following quantities, assuming that when we start observing the chain, it is already in steady-state.



(a) For each state i, the probability that the current state is i.

Problem 2 (Tutorial 9 Problem 3)

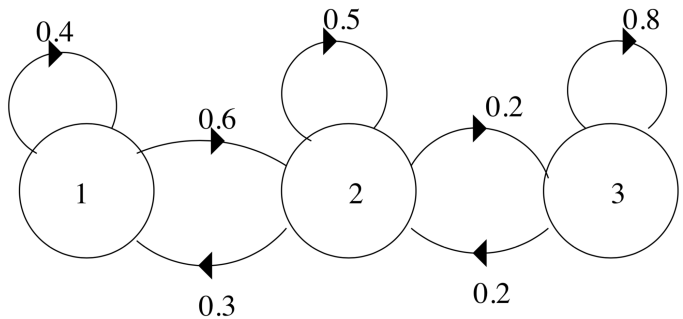
Consider the Markov chain below. Let us refer to a transition that results in a state with a higher (respectively, lower) index as a birth (respectively, death). Calculate the following quantities, assuming that when we start observing the chain, it is already in steady-state.



(a) For each state i, the probability that the current state is i.
The steady state equations take the form $0.6v[1] = 0.3v[2], 0.2v[2] = 0.2v[3]$.
These can be solved, together with the normalization equation, to yield
$v[1] = 1/5, v[2] = v[3] = 2/5$.

Problem 2 (Tutorial 9 Problem 3)

Consider the Markov chain below. Let us refer to a transition that results in a state with a higher (respectively, lower) index as a birth (respectively, death). Calculate the following quantities, assuming that when we start observing the chain, it is already in steady-state.

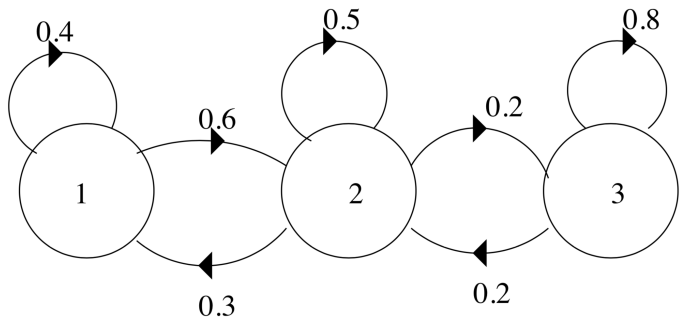


(a) For each state i, the probability that the current state is i.
The steady state equations take the form $0.6v[1] = 0.3v[2], 0.2v[2] = 0.2v[3]$.
These can be solved, together with the normalization equation, to yield
$v[1] = 1/5, v[2] = v[3] = 2/5$.

(b) The probability that the first transition we observe is a birth.

Problem 2 (Tutorial 9 Problem 3)

Consider the Markov chain below. Let us refer to a transition that results in a state with a higher (respectively, lower) index as a birth (respectively, death). Calculate the following quantities, assuming that when we start observing the chain, it is already in steady-state.



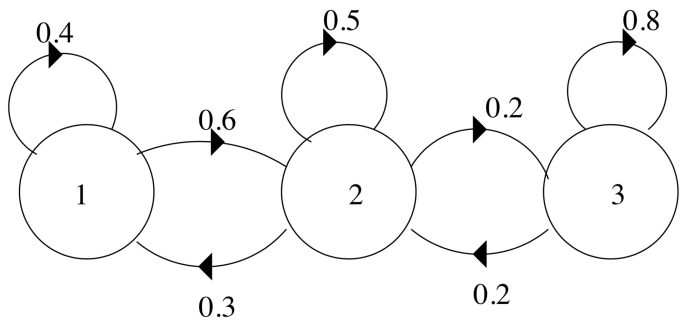
(a) For each state i, the probability that the current state is i.
The steady state equations take the form $0.6v[1] = 0.3v[2], 0.2v[2] = 0.2v[3]$.
These can be solved, together with the normalization equation, to yield
$v[1] = 1/5, v[2] = v[3] = 2/5$.

(b) The probability that the first transition we observe is a birth.

$$0.6v[1] + 0.2v[2] = 0.6\frac{1}{5} + 0.2\frac{2}{5} = \frac{1}{5}$$

Problem 2 (Tutorial 9 Problem 3)

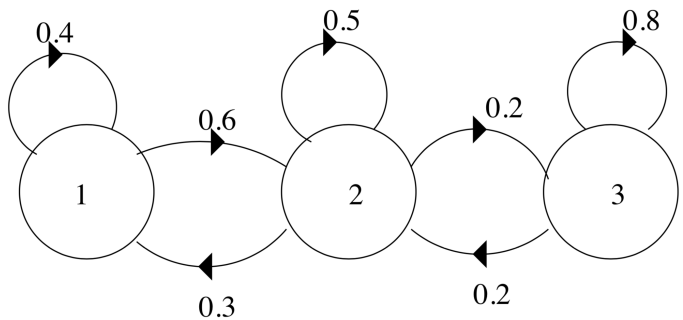
Consider the Markov chain below. Let us refer to a transition that results in a state with a higher (respectively, lower) index as a birth (respectively, death). Calculate the following quantities, assuming that when we start observing the chain, it is already in steady-state.



(c) The probability that the first change of state we observe is a birth.

Problem 2 (Tutorial 9 Problem 3)

Consider the Markov chain below. Let us refer to a transition that results in a state with a higher (respectively, lower) index as a birth (respectively, death). Calculate the following quantities, assuming that when we start observing the chain, it is already in steady-state.

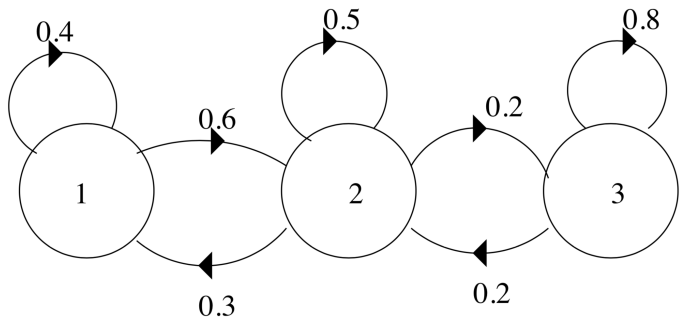


(c) The probability that the first change of state we observe is a birth.
If the state is 1, which happens with probability $1/5$, the first change of state is certain to be a birth. If the state is 2, which happens with probability $2/5$, the probability that the first change of state is a birth is equal to $0.2/(0.3+0.2) = 2/5$. Finally, if the state is 3, the probability that the first change of state is a birth is equal to 0. Thus, the probability that the first change of state that we observe is a birth is equal to

$$1\frac{1}{5} + \frac{2}{5} \cdot \frac{2}{5} = \frac{9}{25}$$

Problem 2 (Tutorial 9 Problem 3)

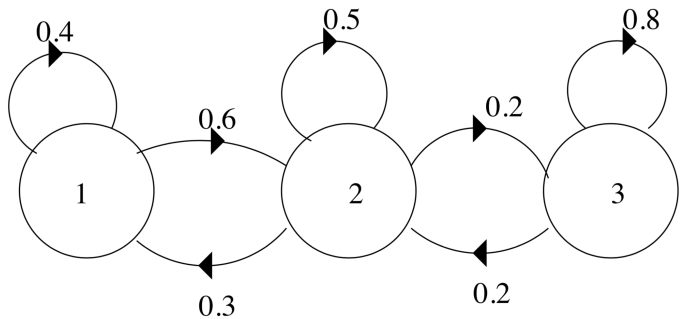
Consider the Markov chain below. Let us refer to a transition that results in a state with a higher (respectively, lower) index as a birth (respectively, death). Calculate the following quantities, assuming that when we start observing the chain, it is already in steady-state.



(d) The conditional probability that the process was in state 2 before the first transition that we observe, given that this transition was a birth.

Problem 2 (Tutorial 9 Problem 3)

Consider the Markov chain below. Let us refer to a transition that results in a state with a higher (respectively, lower) index as a birth (respectively, death). Calculate the following quantities, assuming that when we start observing the chain, it is already in steady-state.

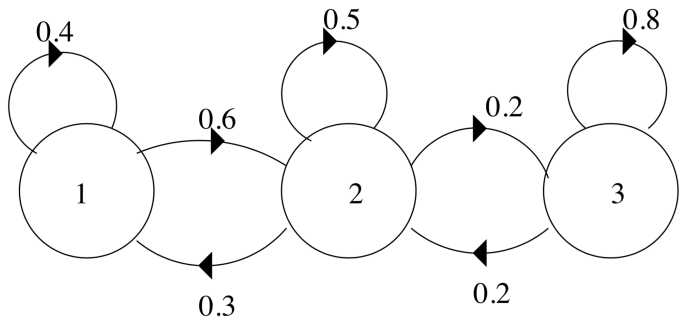


(d) The conditional probability that the process was in state 2 before the first transition that we observe, given that this transition was a birth.

$$P(\text{state was 2} \mid \text{first transition is a birth}) = \frac{\text{state was 2 \& first transition is a birth}}{\text{first transition is a birth}}$$
$$= \frac{0.2v[2]}{0.2} = \frac{2}{5}$$

Problem 2 (Tutorial 9 Problem 3)

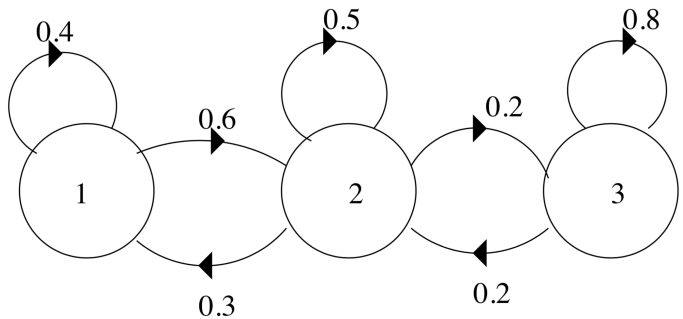
Consider the Markov chain below. Let us refer to a transition that results in a state with a higher (respectively, lower) index as a birth (respectively, death). Calculate the following quantities, assuming that when we start observing the chain, it is already in steady-state.



(e) The conditional probability that the process was in state 2 before the first change of state that we observe, given that this change of state was a birth.

Problem 2 (Tutorial 9 Problem 3)

Consider the Markov chain below. Let us refer to a transition that results in a state with a higher (respectively, lower) index as a birth (respectively, death). Calculate the following quantities, assuming that when we start observing the chain, it is already in steady-state.

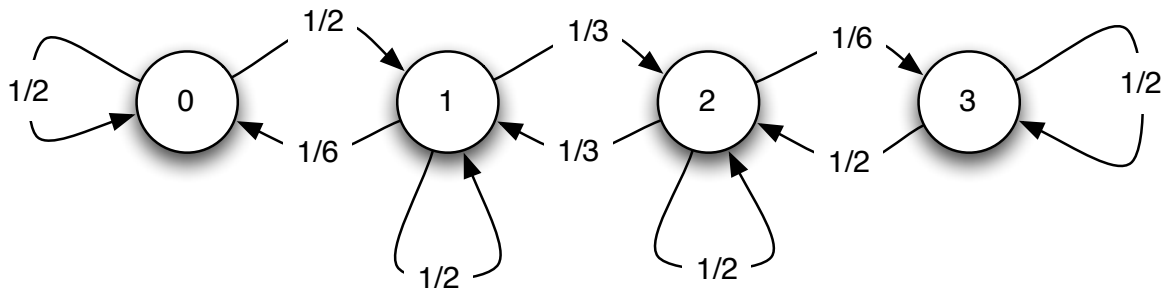


(e) The conditional probability that the process was in state 2 before the first change of state that we observe, given that this change of state was a birth. As shown in part (c), the probability that the first change of state is a birth is $9/25$. Furthermore, the probability that the state is 2 and the first change of state is a birth is $2v[2]/5 = 4/25$. Therefore, the desired probability is

$$\frac{4/25}{9/25} = \frac{4}{9}$$

Lectures' summaries

Markov Chains



Markov Chain

- Markov chains are used for problems in which the future depends on past ONLY through present!
- The condition of the future is summarized by a state, which changes over time according to given probabilities.
- Example: whether you would understand the content of the next class only depends on whether you understand the concept in today's class.

Discrete Markov Chain

- For *discrete*-time Markov chains, the state changes at certain *discrete* time instances, indexed by an integer variable t.
- A discrete Markov chain defines a series of random variables X_t, e.g., $\{X_0, X_1, X_2, \ldots\}$.
- A Markov Chain consists:
  - **State space**: a set of states in which the chain can be described at time t:

$$\mathcal{S} = \{s_1, \ldots, s_k\}$$

  - **Transition probabilities** that describe the probability of transitioning from a state at $t-1$ to another state at t:

$$PX_t = s_j | X_{t-1} = s_i = p_{ij} \text{ for all } 1 \leq i, j \leq k$$

  - **An initial state** X_0, in which the chain is initiated.

Markov Property

- The key assumption is that the transition probabilities (p_{ij}) for the state at time $t+1$ (state j) only depends on the state at time t (state i).
  - The value of X_{t+1} only depends on the value of X_t.
- Mathematically, the **Markov property** defines that

$$P(X_{t+1} = j | X_t = i, X_{t-1} = x_{t-1}, \cdots, X_0 = x_0)$$

$$= P(X_{t+1} = j | X_t = i)$$

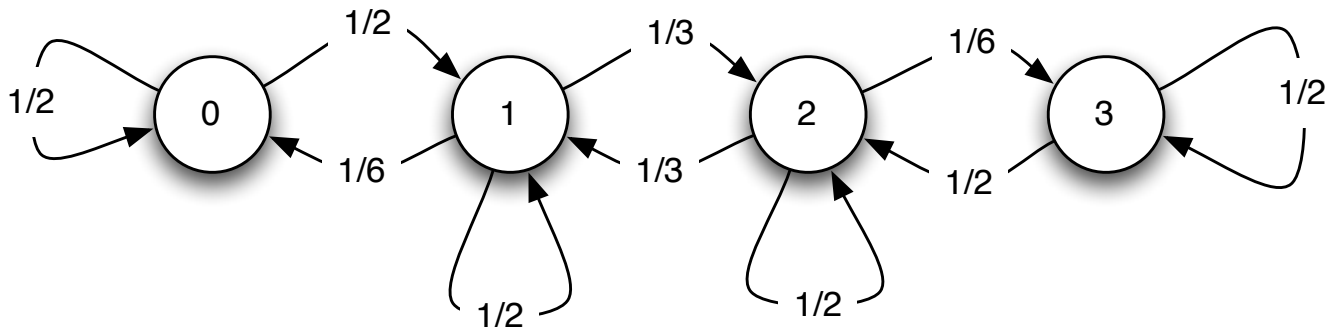
$$= p_{ij}$$

- The transition probability p_{ij} must be **non-negative** and **sum to 1**:

$$\sum_{j=1}^{k} p_{ij} = 1, \text{ for all } i.$$

Transition Probability Graph

A Markov chain can be described using **transition probability graph**, whose nodes are the states and whose arrows are the possible transitions (with probabilities).



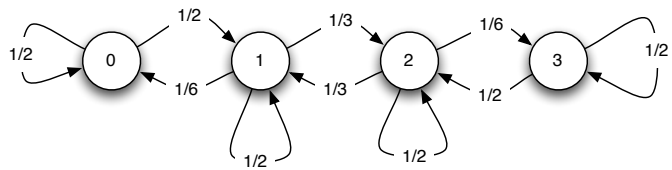
Weights on arrows out of each state i sum to one: $\sum_j p_{ij} = 1$

What if the current state is uncertain?

- What if we don't know X_{t-1}, but know $P(X_{t-1} = i)$ for each i, what's $P(X_t = j)$?
- Then, by the **Law of Total Probability**:

$$\begin{aligned} P(X_t = j) &= \sum_i P(X_t = j, X_{t-1} = i) \\ &= \sum_i P(X_t = j | X_{t-1} = i) P(X_{t-1} = i) \\ &= \sum_i p_{ij} P(X_{t-1} = i) \end{aligned}$$

- **Example**: If there's a $1/3$ probability we're in state 1 and a $2/3$ probability we're in state 3, what's the probability we're in state 2 after one step.



Answer: $4/9$.

Markov Chain Theorem

Theorem

We define the **distribution** *of* X_t *as*

$$\begin{aligned} v_t &= \langle v_t[1], v_t[2], \cdots, v_t[k] \rangle \\ &= \langle PX_t = 1, PX_t = 2, \ldots, PX_t = k \rangle . \end{aligned}$$

where

$$\begin{aligned} v_t[j] &= PX_t = j \\ &= \sum_i PX_t = j | X_{t-1} = i PX_{t-1} = i \\ &= \sum_i p_{ij} v_{t-1}[i]. \end{aligned}$$

Thus,

$$v_t = \left\langle \sum_i p_{i1} v_{t-1}[i], \sum_i p_{i2} v_{t-1}[i], \cdots, \sum_i p_{ik} v_{t-1}[i] \right\rangle .$$

Markov Chain Theorem

This implies that if we know the distribution at $t = 0$ (i.e., v_0), then we can compute any v_t where $t > 0$:

$$v_1 = \left\langle \sum_i p_{i1} v_0[i], \sum_i p_{i2} v_0[i], \cdots, \sum_i p_{ik} v_0[i] \right\rangle .$$

$$v_2 = \left\langle \sum_i p_{i1} v_1[i], \sum_i p_{i2} v_1[i], \cdots, \sum_i p_{ik} v_1[i] \right\rangle .$$

$$v_3 = \left\langle \sum_i p_{i1} v_2[i], \sum_i p_{i2} v_2[i], \cdots, \sum_i p_{ik} v_2[i] \right\rangle .$$

$$\vdots$$

$$v_t = \left\langle \sum_i p_{i1} v_{t-1}[i], \sum_i p_{i2} v_{t-1}[i], \cdots, \sum_i p_{ik} v_{t-1}[i] \right\rangle .$$

Markov Chain Theorem

- This theorem can be **effectively** represented using matrices, but it requires knowledge about linear algebra.
- To give you a short overview, a Markov chain model can be encoded in a **transition probability matrix**. Make sure that you remember the following notation:

$$A = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,k} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k,1} & p_{k,2} & \cdots & p_{k,k} \end{pmatrix}$$

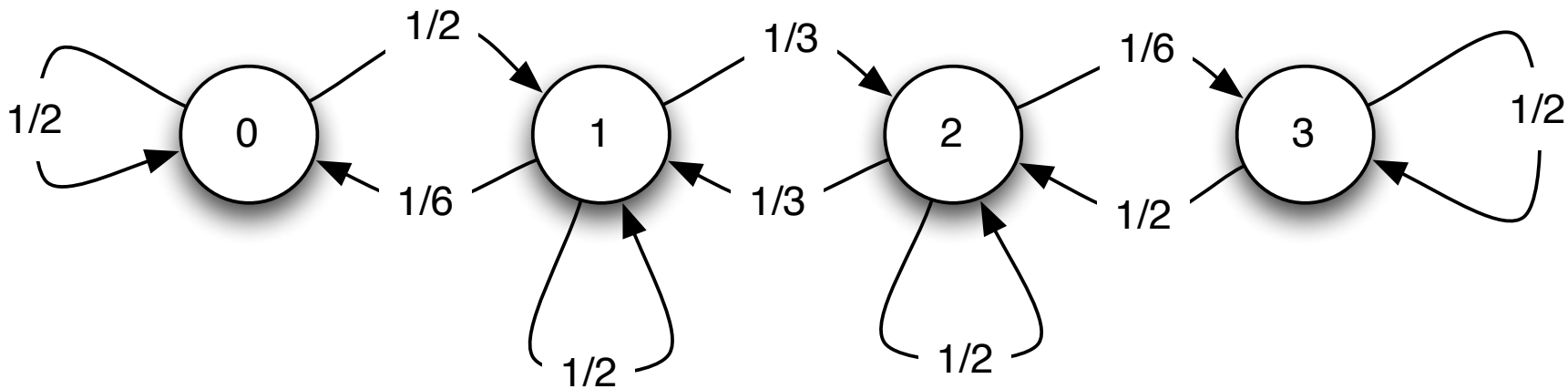
- **Markov Chain Theorem**: Given v_0, we can compute $v_1 = v_0 A$, and

$$v_t = v_{t-1}A = v_{t-2}AA = v_{t-3}AAA = \ldots = v_0 A^t$$

Analyzing Markov Chains via Matrices

- Define **Transition probability matrix**:

$$A = \begin{pmatrix} p_{0,0} & p_{0,1} & p_{0,2} & p_{0,3} \\ p_{1,0} & p_{1,1} & p_{1,2} & p_{1,3} \\ p_{2,0} & p_{2,1} & p_{2,2} & p_{2,3} \\ p_{3,0} & p_{3,1} & p_{3,2} & p_{3,3} \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/6 & 1/2 & 1/3 & 0 \\ 0 & 1/3 & 1/2 & 1/6 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}$$



Simulation of the queue if there is initially one person

Given

$$v_t = \left\langle \sum_i p_{i1} v_{t-1}[i], \sum_i p_{i2} v_{t-1}[i], \cdots, \sum_i p_{ik} v_{t-1}[i] \right\rangle .$$

$$\begin{aligned}
v_0 &= \langle 0.000, 1.000, 0.000, 0.000 \rangle \\
v_1 &= \langle 0.167, 0.500, 0.333, 0.000 \rangle \\
v_2 &= \langle 0.167, 0.444, 0.333, 0.056 \rangle \\
v_3 &= \langle 0.158, 0.416, 0.342, 0.084 \rangle \\
v_4 &= \langle 0.148, 0.401, 0.352, 0.099 \rangle \\
v_5 &= \langle 0.142, 0.391, 0.359, 0.109 \rangle \\
v_6 &= \langle 0.136, 0.386, 0.364, 0.114 \rangle \\
v_7 &= \langle 0.133, 0.382, 0.368, 0.118 \rangle \\
v_8 &= \langle 0.130, 0.380, 0.370, 0.120 \rangle \\
\vdots &\quad \vdots \quad \vdots \\
v_\infty &= \langle 0.125, 0.375, 0.375, 0.125 \rangle
\end{aligned}$$

Steady State Distribution

Do all Markov chains have the property that eventually the distribution settles to the "same steady" state regardless of the initial state?

Definition

We have

$$v = \lim_{t \to \infty} v_t$$

$$\langle v[1], v[2], \ldots, v[k] \rangle = \lim_{t \to \infty} \langle v_t[1], v_t[2], \ldots, v_t[k] \rangle$$

If we have

$$v[j] = \sum_{i=1}^{k} p_{ij} v[i] \text{ for } j = 1, \cdots, k$$

and

$$\sum_{j=1}^{k} v[j] = 1$$

Then, we say v is a **steady state distribution for the Markov Chain.**

Queuing Example

For the queuing example, we had

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} & 0 \\ 0 & \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

if $\mathbf{v} = (v[1], v[2], v[3], v[4])$ then

$$v[1] = \frac{v[1]}{2} + \frac{v[2]}{6}$$

$$v[2] = \frac{v[1]}{2} + \frac{v[2]}{2} + \frac{v[3]}{3}$$

$$v[3] = \frac{v[2]}{3} + \frac{v[3]}{2} + \frac{v[4]}{2}$$

$$v[4] = \frac{v[3]}{6} + \frac{v[4]}{2}$$

Furthermore,

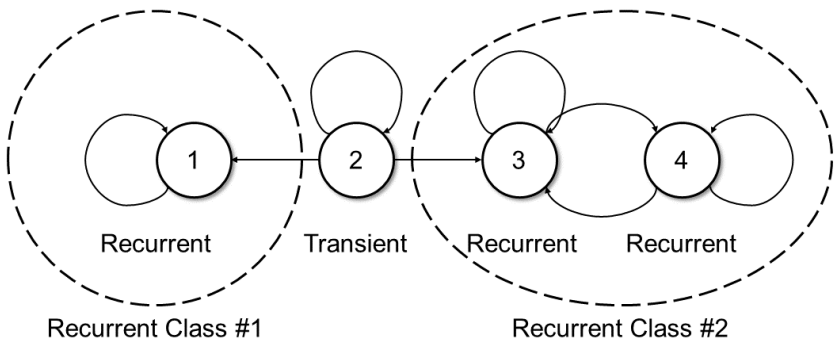
$$v[1] + v[2] + v[3] + v[4] = 1$$

Solving these gives us

$$\mathbf{v} = (0.125, 0.375, 0.375, 0.125)$$

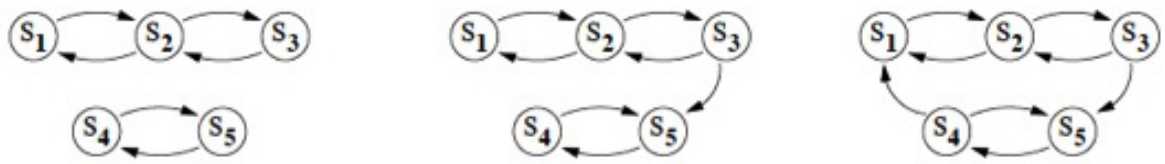
Classification of States

- We say that a state i is **recurrent** if for every j that is accessible from i, i is also accessible from j.
  - Denoting $A(i)$ as a set of states that are accessible from i, for all j that belong to $A(i)$ we have that i belongs to $A(j)$.
- If i is a recurrent state, the set of states $A(i)$ that are accessible from i form a **recurrent class**.
  - States in $A(i)$ are all accessible from each other, and no state outside $A(i)$ is accessible from them.
- A state is called **transient** if it is not recurrent.
- A Markov chain with multiple recurrent classes **does not** converges to a unique steady state.



Recurrent States

- **Question**: Which of the following Markov chains have a single recurrent class?



- **Answer**: Right two chains

Periodic Recurrent Class

Definition

- Consider a recurrent class.
- Let us group all the states into d disjoint groups of states $S_1, \cdots, S_d$; a group has to contain at least one state.
- Such a recurrent class is called **periodic** if there exists at least one group (of states) in the chain that is visited with a period of T. That is, group(s) are visited at time $\{T, 2T, 3T, 4T, \ldots\}$ steps for $T \in \{2, 3, \ldots\}$.
- If a recurrent class is not periodic, we call the class **aperiodic**.

Periodic/Aperiodic Class

- **Question**: Which of the following Markov chains contain a single periodic recurrent class?

- **Answer**: Only the one to the left (with period of 2).

Steady-State Convergence Theorem

Theorem

Consider a Markov chain with a single, aperiodic recurrent class. Then, the states in such a Markov chain have steady-state distribution.

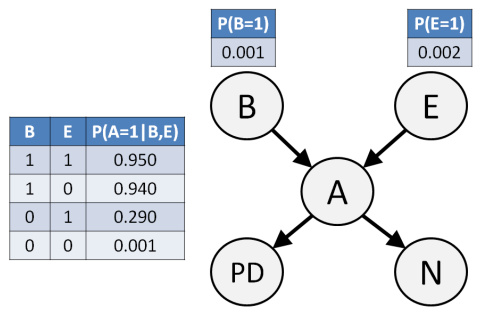
- **Example:** Consider a Markov chain C with 2 states and transition matrix

$$A = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$$

  for some $0 < a, b < 1$
- Does C have a single recurrent class? **Yes.**
- Is C periodic? **No**, as long as $0 < a, b < 1$
- Then, what is its steady state distribution $\mathbf{v}$?
- Let $\mathbf{v} = (c, 1-c)$ be a steady state distribution.
- Solving $v[j] = \sum_{k=1}^{m} v[k]p_{kj}$ for gives:

$$v^* = \left(\frac{b}{a+b}, \frac{a}{a+b} \right)$$

Bayesian Networks



B	E	P(A=1|B,E)
1	1	0.950
1	0	0.940
0	1	0.290
0	0	0.001

A	P(PD=1|A)
1	0.900
0	0.005

A	P(N=1|A)
1	0.750
0	0.100

Chain Rule – Review

- Simplest form of the chain rule is

$$P(A, B) = P(B|A)P(A) = P(A|B)P(B)$$

- Chain rule for 3 variables

$$\begin{aligned} P(A, B, C) &= P(C|A, B)P(A|B)P(B) \\ &= P(C|A, B)P(B|A)P(A) \\ &= P(B|A, C)P(A|C)P(C) \\ &= P(B|A, C)P(C|A)P(A) \\ &= P(A|B, C)P(B|C)P(C) \\ &= P(A|B, C)P(C|B)P(B) \end{aligned}$$

- This can be generalized as

$$P(X_n, \cdots, X_1) = P(X_n|X_{n-1}, \cdots, X_1)P(X_{n-1}, \cdots, X_1)$$

Joint and Marginal Probabilities – Review

- For two discrete random variables X and Y, the joint PMF $P(X, Y)$ was defined as

$$P(X = x, Y = y) = P(X = x \text{ and } Y = y) = P(\{X = x\} \cap \{Y = y\})$$

- Marginal probabilities could be computed as

$$P(X = x) = \sum_{y} P(X = x, Y = y)$$

$$P(Y = y) = \sum_{x} P(X = x, Y = y)$$

- For multiple discrete random variables $X_1, \cdots X_n$ whose joint PMF is denoted as $P(X_1, \cdots X_n)$, marginal probabilities could be computed as

$$P(X_1 = x_1) = \sum_{x_2} \cdots \sum_{x_n} P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$$

Joint PMFs for Many Random Variables

- Before we can think about inference or estimation problems with many random variables, we need to think about the implications of representing joint PMFs over many random variables.
- Why joint PMFs of all random variables?
  - It allows us to compute (marginal or conditional) probabilities of any event that we are interested in.
  - For example, what is the probability that a patient has cancer given test results?

$$P(Cancer|Test_1, \cdots, Test_n) = \frac{P(Cancer, Test_1, \cdots, Test_n)}{P(Test_1, \cdots, Test_n)}$$

The Curse of Dimensionality

- Suppose we have an experiment where we obtain the values of d random variables $X_1, ..., X_d$, where each variable has binary outcomes (for simplicity).
- **Question:** How many **numbers** does it take to write down a joint distribution for them?
- **Answer:** We need to define a probability for each d-bit sequence:

$$P(X_1 = 0, X_2 = 0, ..., X_d = 0)$$
$$P(X_1 = 1, X_2 = 0, ..., X_d = 0)$$
$$\vdots$$
$$P(X_1 = 1, X_2 = 1, ..., X_d = 1)$$

- The number of d-bit sequences is 2^d. Because we know that the probabilities have to add up to 1, we need to write down $2^d - 1$ numbers to specify the full joint PMF on d binary variables.

How Fast is Exponential Growth?

- $2^d - 1$ grows **exponentially** as d increases **linearly**:

d	$2^d - 1$
1	1
10	1023
100	1,267,650,600,228,229,401,496,703,205,375
$\vdots$	$\vdots$

- Storing the full joint PMF for 100 binary variables would take about 10^{30} real numbers or about 10^{18} **terabytes** of storage!
- Joint PMFs grow in size so rapidly, we have no hope whatsoever of storing them explicitly for problems with more than about 30 (binary) random variables.

Factorizing Joint Distributions

- To address this, we start by *factorizing the joint distribution*, i.e., re-writing the joint distribution as a product of conditional PMFs over single variables (called factors).
- If we know some conditional independency between the variables, we can save some space.

Conditional Independence: Simplification Example

- Suppose we instead only assume that:
  - $P(X_2 = a_2|X_1 = a_1, X_3 = a_3) = P(X_2 = a_2|X_1 = a_1)$ for all a_1, a_2, a_3.
- This gives the "conditional independence model": X_2 is conditionally independent of X_3 given X_1

$$\begin{aligned}
&P(X_1 = a_1, X_2 = a_2, X_3 = a_3) \\
&\quad = P(X_1 = a_1)P(X_3 = a_3|X_1 = a_1)P(X_2 = a_2|X_1 = a_1, X_3 = a_3) \\
&\quad = P(X_1 = a_1)P(X_3 = a_3|X_1 = a_1)P(X_2 = a_2|X_1 = a_1)
\end{aligned}$$

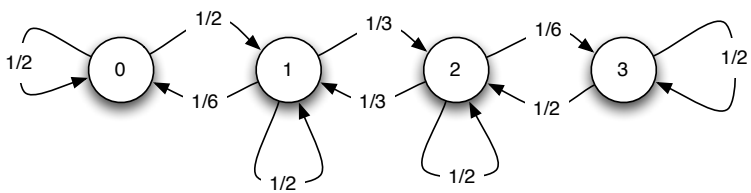
- How many numbers do we need to store for three binary random variables in this case?
  $1 + 2 + 2 = 5$ (as opposed to $2^3 - 1 = 7$ if we encoded the full joint)

Bayesian Networks

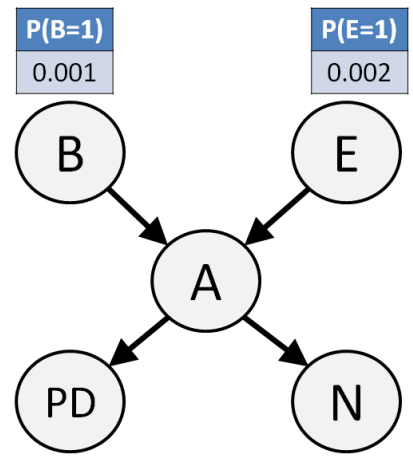
- Keeping track of all the conditional independence assumptions gets tedious when there are a lot of variables.
- To get around this problem, we use "Bayesian Networks" to express the conditional independence structure of these models.
- A Bayesian network uses conditional independence assumptions to more compactly represent a joint PMF of many random variables.
- We use a Directed Acyclic Graph (DAG) to encode conditional independence assumptions.
  - Nodes X_i in the graph G represent random variables.
  - A directed edge $X_j \to X_i$ means X_i directly depends on X_j (not causation!).
  - We also define that X_j is a "parent" of X_i.
  - The set of variables that are parents of X_i is denoted Pa_i.
  - X_i is independent of all its nondescendants given Pa_i.
  - **The factor associated with variable X_i is $P(X_i|Pa_i)$.**

Bayesian Networks vs. Markov Chains

- In Transition Probability Graphs of Markov Chains, **nodes** represent all possible **states**, and **arrows** represents the **probability of transition** from one state to another (with numbers written on it).
- In Bayesian Networks, **nodes** represent all possible **random variables**, and **arrows** represents **dependencies** between the random variables (no numbers associated with it).



B	E	P(A=1|B,E)
1	1	0.950
1	0	0.940
0	1	0.290
0	0	0.001



A	P(PD=1|A)
1	0.900
0	0.005

A	P(N=1|A)
1	0.750
0	0.100

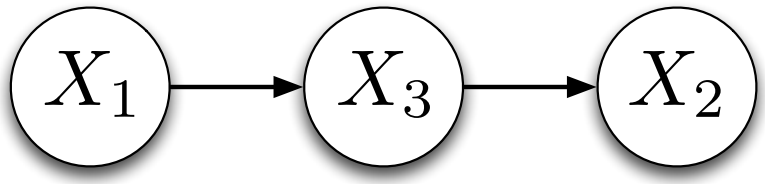
The Bayesian Network Theorem

- **Definition:** A joint PMF $P(X_1, ..., X_d)$ is a Bayesian network with respect to a directed acyclic graph G with parent sets $\{Pa_1, ..., Pa_d\}$ if and only if:

$$P(X_1, ..., X_d) = \prod_{i=1}^{d} P(X_i|Pa_i)$$

- In other words, to be a valid Bayesian network for a given graph G, the joint PMF must factorize according to G.

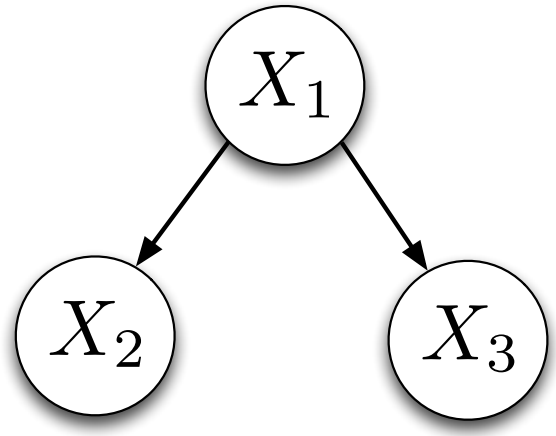
3 Cases of Conditional Independence to Remember



$$Pa_1 = \{\}, Pa_3 = \{X_1\}, Pa_2 = \{X_3\}$$

$$P(X_1, X_2, X_3) = P(X_1)P(X_3|X_1)P(X_2|X_3)$$

3 Cases of Conditional Independence to Remember



$$Pa_1 = \{\}, Pa_3 = \{X_1\}, Pa_2 = \{X_1\}$$

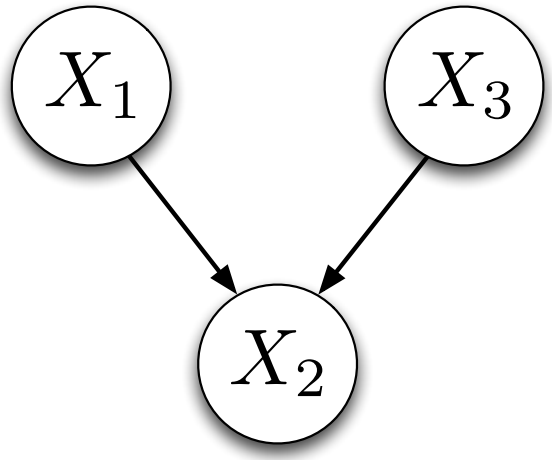
$$P(X_1, X_2, X_3) = P(X_1)P(X_3|X_1)P(X_2|X_1) \quad (1)$$

- Note that X_2 and X_3 are conditionally independent given X_1:

$$P(X_2, X_3|X_1) = P(X_2|X_1) \cdot P(X_3|X_1)$$

Proof: divide both sides in (1) by $P(X_1)$

3 Cases of Conditional Independence to Remember



$$Pa_1 = \{\}, Pa_3 = \{\}, Pa_2 = \{X_1, X_3\}$$

$$P(X_1, X_2, X_3) = P(X_1)P(X_3)P(X_2|X_1, X_3) \quad (2)$$

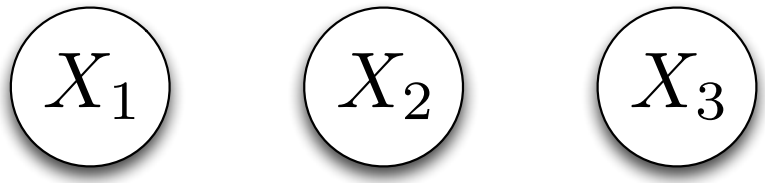
- Note that X_1 is not independent of X_3 given X_2:

$$P(X_1, X_3|X_2) \neq P(X_1|X_2) \cdot P(X_3|X_2)$$

Proof: divide both sides in (2) by $P(X_2)$:

$$P(X_1, X_3|X_2) = \frac{P(X_1)P(X_3)P(X_2|X_1, X_3)}{P(X_2)} \neq P(X_1|X_2) \cdot P(X_3|X_2)$$

If All Nodes Are Independent

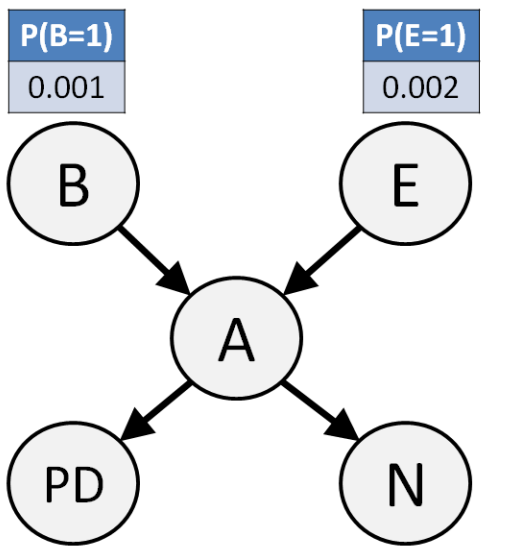


$$Pa_1 = \{\}, Pa_3 = \{\}, Pa_2 = \{\}$$

$$P(X_1, X_2, X_3) = P(X_1)P(X_3)P(X_2)$$

The Alarm Network: Random Variable

- You live in quiet neighborhood in the suburbs of LA. There are two reasons the alarm system in your house will go off: your house is broken into or there is an earthquake. If your alarm goes off you might get a call from the police department. You might also get a call from your neighbor.



B	E	P(A=1|B,E)
1	1	0.950
1	0	0.940
0	1	0.290
0	0	0.001

A	P(PD=1|A)
1	0.900
0	0.005

A	P(N=1|A)
1	0.750
0	0.100

$$P(B, E, A, PD, N) = P(B)P(E)P(A|B, E)P(PD|A)P(N|A)$$

The Alarm Network: Marginal Query

- **Question:** What is the probability that there was a break-in, but no earthquake, the police call, but your neighbor does not call?

$$
\begin{aligned}
&P(B=1,E=0,PD=1,N=0)\\
&=\sum_{A=\{0,1\}} P(B=1,E=0,PD=1,N=0,A)\\
&=P(B=1,E=0,PD=1,N=0,A=0)\\
&+P(B=1,E=0,PD=1,N=0,A=1)\\
&=P(B=1)P(E=0)P(A=1|B=1,E=0)P(PD=1|A=1)P(N=0|A=1)\\
&+P(B=1)P(E=0)P(A=0|B=1,E=0)P(PD=1|A=0)P(N=0|A=0)\\
&\\
&=0.001\cdot(1-0.002)\cdot 0.94\cdot 0.9\cdot(1-0.75)\\
&+0.001\cdot(1-0.002)\cdot(1-0.94)\cdot 0.005\cdot(1-0.1)=0.00021\ldots
\end{aligned}
$$

The Alarm Network: Marginal Query

- **Question:** What is the probability that the alarm will be on?

$$
\begin{aligned}
& P(A=1) \\
&= \sum_{B}\sum_{E}\sum_{PD}\sum_{N} P(A=1,B,E,PD,N) \\
&= \sum_{B}\sum_{E}\sum_{PD}\sum_{N} P(B)P(E)P(A=1|B,E)P(PD|A=1)P(N|A=1) \\
&= P(B=0)P(E=0)P(A=1|B,E=0)P(PD=0|A=1)P(N=0|A=1) \\
&+P(B=0)P(E=0)P(A=1|B,E=0)P(PD=0|A=1)P(N=1|A=1) \\
&\cdots \\
&+P(B=1)P(E=1)P(A=1|B,E=1)P(PD=1|A=1)P(N=1|A=1)
\end{aligned}
$$

The Alarm Network: Marginal Query

- We can compute the above using a simple algorithm:

$Z = 0;$
for $B = 0$ *to* 1 **do**
  for $E = 0$ *to* 1 **do**
    for $PD = 0$ *to* 1 **do**
      for $N = 0$ *to* 1 **do**
        $Z = Z +$
          $P(B)P(E)P(A = 1|B, E)P(PD|A = 1)P(N|A = 1);$
      end
    end
  end
end

- What would be the potential problem with this?
  - Computational complexity explodes as # of variables increases
  - The multiplication of small number approaches to 0 as # of variables increases

The Alarm Network: Marginal Query

- We can optimize the computation as the following

$$
\begin{aligned}
& P(A=1) \\
&= \sum_{B}\sum_{E}\sum_{PD}\sum_{N} P(B)P(E)P(A=1|B,E)P(PD|A=1)P(N|A=1) \\
&= \sum_{B}\sum_{E} P(B)P(E)P(A=1|B,E)\sum_{PD} P(PD|A=1)\sum_{N} P(N|A=1) \\
&= \sum_{B}\sum_{E} P(B)P(E)P(A=1|B,E)
\end{aligned}
$$

The Alarm Network: Conditional Query

- **Question:** What is the probability that the alarm went off **given that** there was a break-in, but no earthquake, the police call, but your neighbor does not call?

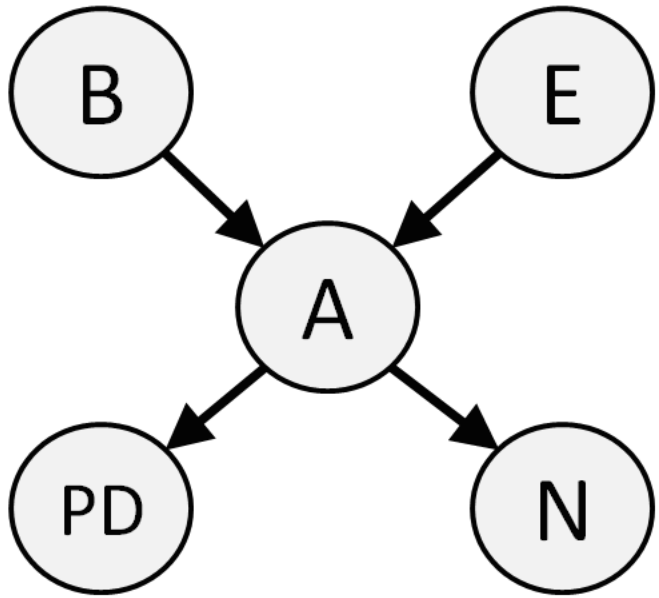
$$
\begin{aligned}
& P(A = 1 | B = 1, E = 0, PD = 1, N = 0) \\
&= \frac{P(B = 1, E = 0, A = 1, PD = 1, N = 0)}{P(B = 1, E = 0, PD = 1, N = 0)} \\
&= \frac{P(B = 1, E = 0, A = 1, PD = 1, N = 0)}{\sum_{a=0}^{1} P(B = 1, E = 0, A = a, PD = 1, N = 0)} \\
&= \frac{P(B = 1)P(E = 0)P(A = 1 | B = 1, E = 0)P(PD = 1 | A = 1)P(N = 0 | A = 1)}{\sum_{a=0}^{1} P(B = 1)P(E = 0)P(A = a | B = 1, E = 0)P(PD = 1 | A = a)P(N = 0 | A = a)}
\end{aligned}
$$

Answering Probabilistic Queries

- **Joint Query:** To compute the probability of an assignment to all of the variables we simply express the joint probability as a product over the individual factors. We then look up the correct entries in the factor tables and multiply them together.
- **Marginal Query:** To compute the probability of an observed subset of the variables in the Bayesian network, we sum the joint probability of all the variables over the possible configurations of the unobserved variables.
- **Conditional Query:** To compute the probability of one subset of the variables given another subset, we first apply the conditional probability formula and then compute the ratio of the resulting marginal probabilities.

Estimating Bayesian Networks from Data

- Just as with simpler models like the biased coin, we can estimate the unknown model parameters from data.



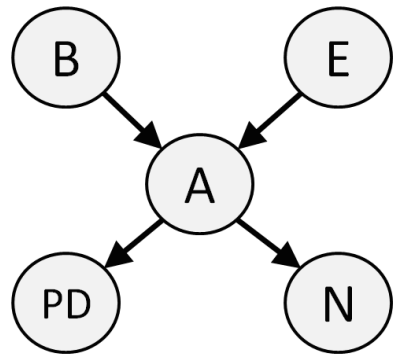
- If we have data consisting of n observations of all of the variables in the network, we can easily estimate the entries of each conditional probability table.

Estimating Bayesian Networks: Counting

- **No Parents:** For a variable X with no parents, the estimate of $P(X = x)$ is just the number of times that the variable X takes the value x in the data, divided by the total number of data cases n.
- **Some Parents:** For a variable X with parents $Y_1, ..., Y_p$, the estimate of $P(X = x | Y_1 = y_1, ..., Y_p = y_p)$ is just the number of times that the variable X takes the value x when the parent variables $Y_1, ..., Y_p$ take the values $y_1, ..., y_p$, divided by the total number of times that the parent variables take the values $y_1, ..., y_p$.

Computing the Factor Tables from Observations

- Suppose we have a sample of data as shown below. Each row i is a joint configuration of all of the random variables in the network.



E	B	A	PD	N
1	0	1	1	1
0	0	0	0	1
0	0	1	1	0
0	1	1	1	0
0	0	0	0	0

- In the alarm network, consider the factor $P(E)$. We need to estimate $P(E = 0)$ and $P(E = 1)$.
- Given our data sample, we get the answers $P(E = 0) = 4/5$ and $P(E = 1) = 1/5$.

Computing the Factor Tables from Observations

- In the alarm network, consider the factor $P(N|A)$. We need to estimate $P(N = 0|A = 0)$,$P(N = 1|A = 0)$, $P(N = 0|A = 1)$, $P(N = 1|A = 1)$. How can we do this?

E	B	A	PD	N
1	0	1	1	1
0	0	0	0	1
0	0	1	1	0
0	1	1	1	0
0	0	0	0	0

- $P(N = 0|A = 0) = \frac{1}{2}, P(N = 1|A = 0) = \frac{1}{2}$
- $P(N = 0|A = 1) = \frac{2}{3}, P(N = 1|A = 1) = \frac{1}{3}$