# Uncertainty & Probabilistic Reasoning

10S3001 – Artificial Intelligence

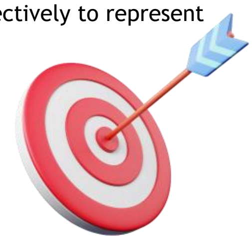
Samuel I. G. Situmeang

Faculty of Informatics and Electrical Engineering

# Objectives

Students are able:

- to explain the concept of probabilistic reasoning intuitively and mathematically, including a basic understanding of Bayes' theorem and its application in decision making under uncertainty.
- to select and use probabilistic models based on Bayes' theorem effectively to represent and analyze uncertainty in various situations such as classification, diagnosis, or prediction.
- to apply Markov Chain and Hidden Markov Models for time series analysis and pattern recognition, and evaluate the performance of the resulting models.





- Mahasiswa mampu menjelaskan konsep bernalar probabilistik secara intuitif dan matematis, termasuk pemahaman dasar teorema Bayes serta penerapannya dalam pengambilan keputusan di bawah ketidakpastian.
- Mahasiswa mampu memilih dan menggunakan model probabilistik berbasis teorema Bayes secara efektif untuk merepresentasikan dan menganalisis ketidakpastian dalam berbagai situasi seperti klasifikasi, diagnosa, atau prediksi.
- Mahasiswa mampu menerapkan *Markov Chain* dan *Hidden Markov Model* untuk analisis deret waktu dan pengenalan pola, serta mengevaluasi performa model yang dihasilkan.

3

# Reasoning Under Uncertainty

- Why do we need reasoning under uncertainty?
- How should uncertainty be represented?

# Why do we need reasoning under uncertainty?

- There are many situations where uncertainty arises:
  - When you travel you reason about the possibility of delays
  - When an insurance company offers a policy it has calculated the risk that you will claim
  - When your brain estimates what an object is it filters random noise and fills in missing details
  - When you play a game you cannot be certain what the other player will do
  - A medical expert system that diagnoses disease has to deal with the results of tests that are sometimes incorrect
- Systems which can reason about the effects of uncertainty should do better than those that don't
- But how should uncertainty be represented?

# Two (toy) examples

- I have toothache. What is the cause?

  There are many possible causes of an observed event.

- If I go to the dentist and he examines me, when the probe catches this indicates there may be a cavity, rather than another cause.

  The likelihood of a hypothesised cause will change as additional pieces of evidence arrive.

- Bob lives in San Francisco. He has a burglar alarm on his house, which can be triggered by burglars and earthquakes. He has two neighbours, John and Mary, who will call him if the alarm goes off while he is at work, but each is unreliable in their own way. All these sources of uncertainty can be quantified. Mary calls, how likely is it that there has been a burglary?

  Using probabilistic reasoning we can calculate how likely a hypothesised cause is.

# Probability Theory: Variables and Events

- A random variable can be an observation, outcome or event the value of which is uncertain.
  - e.g. A coin. Let's use Throw as the random variable denoting the outcome when we toss the coin.
- The set of possible outcomes for a random variable is called its domain.
  - e.g. A coin. The domain of Throw is {head, tail}
- A Boolean random variable has two outcomes.
  - e.g. Cavity has the domain {true, false}
  - e.g. Toothache has the domain {true, false}

# Probability Theory: Atomic Events

- We can create new events out of combinations of the outcomes of random variables.

- An atomic event is a complete specification of the values of the random variables of interest.
  - e.g. if our world consists of only two Boolean random variables, then the world has a **four** possible atomic events
    - $\text{Toothache} = \text{true} \quad \wedge \quad \text{Cavity} = \text{true}$
    - $\text{Toothache} = \text{true} \quad \wedge \quad \text{Cavity} = \text{false}$
    - $\text{Toothache} = \text{false} \quad \wedge \quad \text{Cavity} = \text{true}$
    - $\text{Toothache} = \text{false} \quad \wedge \quad \text{Cavity} = \text{false}$
  - The set of all possible atomic events has two properties:
    - It is mutually exhaustive (nothing else can happen)
    - It is mutually exclusive (only one of the four can happen at one time)

# Probability Theory: Probabilities

- We can assign probabilities to the outcomes of a random variable.
  - $P(\text{Throw} = \text{heads}) = 0.5$
  - $P(\text{Mary\_Calls} = \text{true}) = 0.1$
  - $P(a) = 0.3$

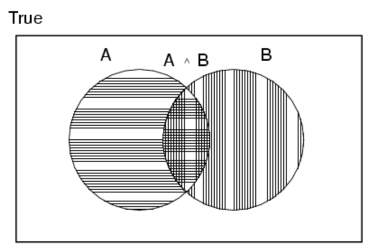
- Some simple rules governing probabilities
  1. All probabilities are between 0 and 1 inclusive: $0 \leq P(a) \leq 1$
  2. If something is necessarily true it has probability 1: $P(\text{true}) = 1$
  3. The probability of a disjunction being true is: $P(\text{false}) = 0$

- From these three laws all of probability theory can be derived.

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

# Probability Theory: Relation to Set Theory

- We can often intuitively understand the laws of probability by thinking about sets




$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

# Probability Theory: Conditional Probability

- A conditional probability expresses the likelihood that one event $a$ will occur if $b$ occurs. We denote this as $P(a|b)$.
  - e.g.

$$P(\text{Toothache} = \text{true}) = 0.2$$
$$P(\text{Toothache} = \text{true}|\text{Cavity} = \text{true}) = 0.6$$

- So, conditional probabilities reflect the fact that some events make other events more (or less) likely.
- If one event doesn't affect the likelihood of another event, they are said to be **independent** and therefore:

$$P(a|b) = P(a)$$

  - e.g. if you roll a die, it doesn't make it more or less likely that you will roll a 6 on the next throw. The rolls are **independent**.

# Combining Probabilities: The Product Rule

- How we can work out the likelihood of two events occuring together given their base and conditional probabilities?

$$P(a \wedge b) = P(a \mid b)P(b) = P(b \mid a)P(a)$$

- So in our toy example 1:

$$\begin{aligned} P(toothache \wedge cavity) &= P(toothache \mid cavity)P(cavity) \\ &= P(cavity \mid toothache)P(toothache) \end{aligned}$$

- But this doesn't help us answer our question:

"I have toothache. Do I have a cavity?"

# Bayes' rule

- We can rearrange the two parts of the product rule:

$$P(a \wedge b) = P(a \mid b)P(b) = P(b \mid a)P(a)$$

$$P(a \mid b)P(b) = P(b \mid a)P(a)$$

$$P(a \mid b) = \frac{P(b \mid a)P(a)}{P(b)}$$

- This is known as **Bayes' rule**.
- It is the cornerstone of modern probabilistic AI.
- But why is it useful?

# Bayes' rule

- We can think about some events as being "hidden" causes: not necessarily directly observed (e.g. a cavity).

- If we model how likely observable effects are given hidden causes (how likely toothache is given a cavity), then Bayes' rule allows us to use that model to infer the likelihood of the hidden cause (and thus answer our question)

$$P(cause \mid effect) = \frac{P(effect \mid cause)P(cause)}{P(effect)}$$

- In fact, good models of $P(\text{effect}|\text{cause})$ are often available to us in real domains (e.g. medical diagnosis)

# Bayes' rule can capture causal models

- Suppose a doctor knows that a meningitis causes a stiff neck in 50% of cases

$$P(s \mid m) = 0.5$$

- She also knows that the probability in the general population of someone having a stiff neck at any time is 1/20

$$P(s) = 0.05$$

- She also has to know the incidence of meningitis in the population (1/50,000)

$$P(m) = 0.00002$$

- Using Bayes' rule she can calculate the probability the patient has meningitis:

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.5 \times 0.00002}{0.05} = 0.0002 = 1/5000$$

$$P(cause \mid effect) = \frac{P(effect \mid cause)P(cause)}{P(effect)}$$

# The power of causal models

- Why wouldn't the doctor be better off if she just knew the likelihood of meningitis given a stiff neck? i.e. information in the diagnostic direction from symptoms to causes?
  - Because diagnostic knowledge is often more fragile than causal knowledge

- Suppose there was a meningitis epidemic? The rate of meningitis goes up 20 times within a group

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.5 \times \boxed{0.0004}}{0.05} = 0.004 = 1/250$$

- The causal model $P(s|m)$ is unaffected by the change in $P(m)$, whereas the diagnostic model $P(m|s) = 1/5000$ is now badly wrong.



- $P(m)_{new} = 20 \times P(m)_{old}$
- $P(m)_{new} = 20 \times 0.00002$

# Bayes' rule: the normalisation short cut

- If we know $P(\text{effect}|\text{cause})$, for every cause we can avoid having to know $P(\text{effect})$

$$P(c \mid e) = \frac{P(e \mid c)P(c)}{P(e)} = \frac{P(e \mid c)P(c)}{\sum_{\forall h \in Causes} P(e \mid h)P(h)}$$

- Suppose for two possible causes of a stiff neck, meningitis ($m$) and not meningitis ($\neg m$)

$$\mathbf{P}(Meningitis) = \alpha < P(s \mid m)P(m), P(s \mid \neg m)P(\neg m) >$$

- We simply calculate the top line for each one and then normalise (divide by the sum of the top line for all hypothesises.
- But sometimes it's harder to find out $P(\text{effect}|\text{cause})$ for all causes independently than it is simply to find out $P(\text{effect})$.
- Note that Bayes' rule here relies on the fact the effect must have arisen because of one of the hypothesised causes. You can't reason directly about causes you haven't imagined.



Denominator can be viewed as a normalization constant α

# Bayes' rule: combining evidence

- Suppose we have several pieces of evidence we want to combine:
  - John rings and Mary rings
  - I have toothache and the dental probe catches on my tooth

- How do we do this?

$$P(cavity \mid toothache \wedge catch) = \alpha\, P(toothache \wedge catch \mid cavity) P(cavity)$$

- As we have more effects our causal model becomes very complicated (for N binary effects there will be $2^N$ different combinations of evidence that we need to model given a cause)

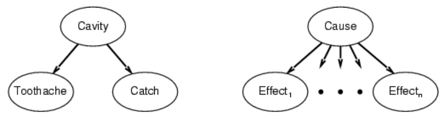
$$P(toothache \wedge catch \mid cavity)\ ,\ \ P(toothache \wedge \neg catch \mid cavity)\ \ \cdots$$

# Bayes' rule + conditional independence

- In many practical applications there are not a few evidence variables but hundreds
- Thus $2^N$ is very big
- This nearly led everyone to give up and rely on approximate or qualitative methods for reasoning about uncertainty
- But conditional independence helps
- Toothache and catch are not independent, but they are independent given the presence or absence of a cavity.
- In other words we can use the knowledge that cavities cause toothache and they cause the catch, but the catch and the toothache do not cause each other (they have a single common cause).

# Bayes' nets

- This can be captured in a picture, where the arcs capture conditional independence relationships



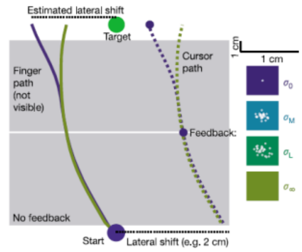

- Or in a new equation:

$$P(cavity \mid toothache \wedge catch) = \alpha\, P(toothache \wedge catch \mid cavity) P(cavity)$$
$$= \alpha\, P(toothache \mid cavity) P(catch \mid cavity) P(cavity)$$

- Using conditional independence the causal model is much more compact. Rather than the number of parameters being $O(2^N)$ it is $O(N)$ where N is the number of effects (or evidence variables)

# The Bayesian brain

- It turns out that your brain understands Bayes rule!




Reproduced without permission from Kording & Wolpert Nature Vol 427, 15 Jan 2004

- The human has to move their (hidden) hand to a target. Half way through the movement they are given an uncertain clue as to where the hand is. The position their hand moves to by the end is consistent with the brain using Bayes' rule to combine the information it receives

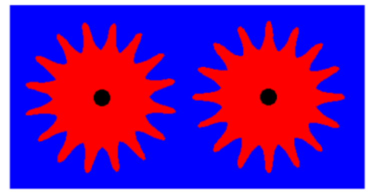
https://mitpress.mit.edu/books/bayesian-brain

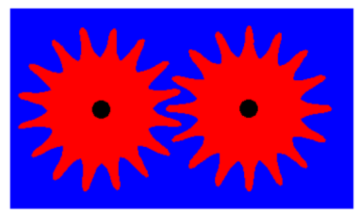


***Bayesian brain*** menyatakan bahwa otak merepresentasikan informasi sensorik secara probabilistik, dalam bentuk distribusi probabilitas.

# The non-Bayesian brain

- But it is not apparent that Bayes’ rule is used everywhere
- How will rotation of one wheel affect the other?
- Some kinds of inference don’t seem to be obviously explainable using probabilistic reasoning alone





# Probabilistic Reasoning Over Time

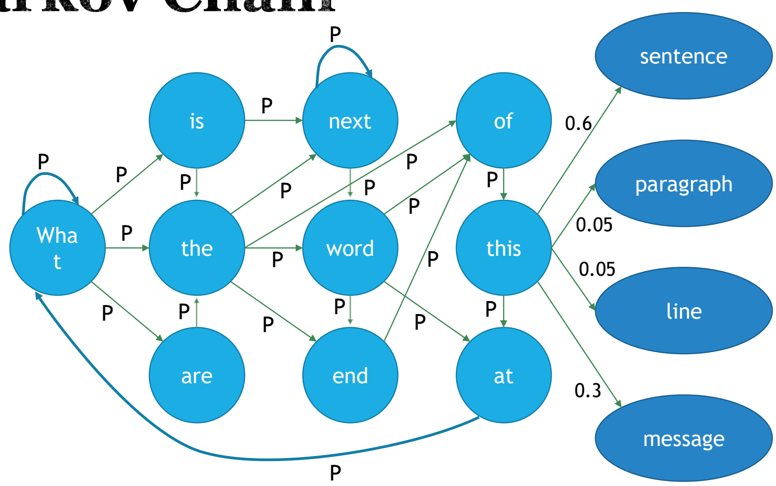
22

- Markov Chain
- Hidden Markov Model

# Motivation

What is the word at the end of this ________?

# Markov Chain

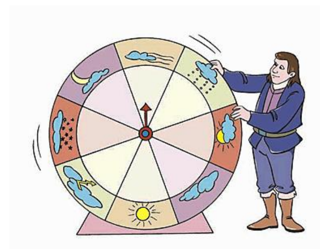


# Markov Chain: Weather Prediction Example

- Design a Markov Chain to predict the weather of tomorrow using previous information of the past days.

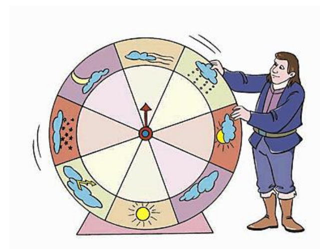
- We have three types of weather: $sunny$, $rainy$, and $cloudy$.
  - So, our model has 3 states: $S = \{S_1, S_2, S_3\}$, and the name of each state is $S_1 = Sunny$, $S_2 = Rainy$, $S_3 = Cloudy$.

- Assume that the weather lasts all day, i.e. it doesn't change from rainy to sunny in the middle of the day.



# Markov Chain: Weather Prediction Example

- Assume a simplified model of weather prediction:
  - Collect statistics on what the weather was like today based on what the weather was like yesterday, the day before, and so forth to collect the following probabilities:
  - $P(w_n|w_{n-1}, w_{n-2}, \dots, w_1)$
  - With the expression, we can give probabilities of types of weather for tomorrow and the next day using $n$ days history.



# Markov Chain: Weather Prediction Example

- For example, the past three days was $\{sunny, sunny, cloudy\}$, the probability that tomorrow would be rainy is given by:
  - $P(w_4 = Rainy | w_3 = Cloudy, w_2 = Sunny, w_1 = Sunny)$

- Problem: the larger $n$ is, the more statistics we must collect. Suppose $n = 5$, then we must collect statistics for $3^5 = 243$ past histories. Therefore, *Markov Assumption*:
  - In a sequence $\{w_1, w_2, \ldots, w_n\}$:

$$P(w_n | w_{n-1}, w_{n-2}, \ldots, w_1) \approx P(w_n | w_{n-1})$$

# Markov Assumption

- $P(w_n|w_{n-1}, w_{n-2}, \dots, w_1) \approx P(w_n|w_{n-1})$ called a *first-order* Markov Assumption, since we say that the probability of an observation at time $n$ only depends on the observation at time $n-1$.
- A *second-order* Markov assumption would have the observation at time $n$ depend on $n-1$ and $n-2$.
- We can express the joint probability using the Markov assumption.

$$P(w_1, \dots, w_n) = \prod_{i=1}^{n} P(w_i|w_{i-1})$$



- One question that comes to mind is "What is $w_0$?" In general, one can think of $w_0$ as the START word, so $\mathrm{P}(w_1|w_0)$ is the probability that $w_1$ can start a sentence. We can also just multiply the prior probability of $w_1$ with the product of $\prod_{i=1}^{n} P(w_i|w_{i-1})$; it's just a matter of definitions.

# Markov Chain: Weather Prediction Example

- Let's arbitrarily pick some numbers for $P(w_{tomorrow}|w_{today})$ expressed in Table 1.

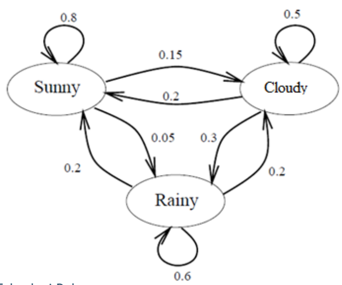
Table 1. Probabilities of Tomorrow's weather based on Today's weather

| | | Tomorrow's Weather | | |
|---|---|---|---|---|
| | | Sunny | Rainy | Cloudy |
| Today's Weather | Sunny | 0.8 | 0.05 | 0.15 |
| | Rainy | 0.2 | 0.6 | 0.2 |
| | Foggy | 0.2 | 0.3 | 0.5 |

- For first-order Markov models, we can use these probabilities to draw a probabilistic finite state automaton.

# Markov Chain: Weather Prediction Example

- For the weather domain, you would have three states (Sunny, Rainy, Cloudy) and every day you would transition to a (possibly) new state based on the probabilities in Table 1.
- Such an automaton would look like this:



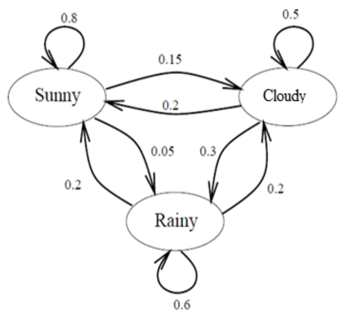

$$\left.\begin{array}{l} P(Sunny|Sunny) = 0.8 \\ P(Rainy|Sunny) = 0.05 \\ P(Cloudy|Sunny) = 0.15 \end{array}\right\} 1$$

$$\left.\begin{array}{l} P(Sunny|Rainy) = 0.2 \\ P(Rainy|Rainy) = 0.6 \\ P(Cloudy|Rainy) = 0.2 \end{array}\right\} 1$$

$$\left.\begin{array}{l} P(Sunny|Cloudy) = 0.2 \\ P(Rainy|Cloudy) = 0.3 \\ P(Cloudy|Cloudy) = 0.5 \end{array}\right\} 1$$

# Markov Chain: Weather Prediction Example

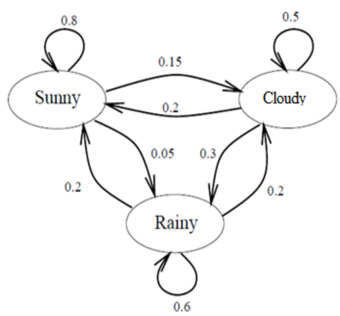
- Exercise 1
  - Given that today is Sunny, what's the probability that tomorrow is Sunny and the day after is Rainy?




$$P(w_2, w_3|w_1) = P(w_3|w_2, w_1) * P(w_2|w_1)$$
$$= P(w_3|w_2\,) * P(w_2|w_1)$$
$$= P(Rainy|Sunny)P(Sunny|Sunny)$$
$$= (0.05)(0.8)$$
$$= 0.04$$

# Markov Chain: Weather Prediction Example

- Exercise 2
  - Given that today is Cloudy, what's the probability that it will be Rainy two days from now?




There are three ways to get from Cloudy today to Rainy two days from now: {Cloudy, Cloudy, Rainy}, {Cloudy, Rainy, Rainy}, and {Cloudy, Sunny, Rainy.}

$$
\begin{aligned}
P(w_3|w_1) = {} & P(w_2 = C, w_3 = R|w_1 = C) + \\
& P(w_2 = R, w_3 = R|w_1 = C) + \\
& P(w_2 = S, w_3 = R|w_1 = C) \\
= {} & P(w_3 = R|w_2 = C) * P(w_2 = C|w_1 = C) + \\
& P(w_3 = R|w_2 = R) * P(w_2 = R|w_1 = C) + \\
& P(w_3 = R|w_2 = S) * P(w_2 = S|w_1 = C) \\
= {} & (0.3)(0.5) + (0.6)(0.3) + (0.05)(0.2) \\
= {} & 0.34
\end{aligned}
$$

# What is A Markov Model?

- A Markov Model is a stochastic model which models temporal or sequential data, i.e., data that are ordered.
- It provides a way to model the dependencies of current information (e.g. weather) with previous information.
- It is composed of states, transition scheme between states, and emission of outputs (discrete or continuous).
- Several goals can be accomplished by using Markov models:
  - Learn statistics of sequential data.
  - Do prediction or estimation.
  - Recognize patterns.



Stochastic: randomly determined; having a random probability distribution or pattern that may be analyzed statistically but may not be predicted precisely.

# Motivation

What makes a Hidden Markov Model?

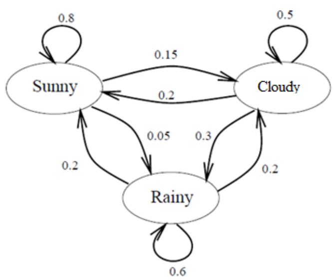
# What is A Hidden Markov Model (HMM)?

- A Hidden Markov Model, is a stochastic model where the states of the model are hidden. Each state can emit an output which is observed.

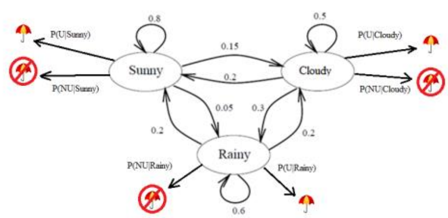
- **Imagine:** You were locked in a room for several days, and you were asked about the weather outside. The only piece of evidence you have is whether the person who comes into the room carrying your daily meal is carrying an umbrella or not.
  - **What is hidden?** Sunny, Rainy, Cloudy
  - **What can you observe?** Umbrella or Not

# Markov Chain vs. Hidden Markov Model

## Markov Chain




## Hidden Markov Model




U = Umbrella
NU= Not Umbrella

# Weather Model

- Let's assume that $t$ days had passed. Therefore, we will have an observation sequence $O = \{o_1, \dots, o_t\}$, where $o_i \in \{Umbrella, Not\ Umbrella\}$.

- Each observation comes from an unknown state. Therefore, we will also have an unknown sequence $W = \{w_1, \dots, w_t\}$, where $w_i \in \{Sunny, Rainy, Cloudy\}$.

- We would like to know: $P(w_1, \dots, w_t | o_1, \dots, o_t)$.

# HMM Mathematical Model

- From Bayes' Theorem, we can obtain the probability for a particular day as:

$$P(w_i|o_i) = \frac{P(o_i|w_i)P(w_i)}{P(o_i)}$$

- For a sequence of length $t$:

$$P(w_1, \ldots, w_t|o_1, \ldots, o_t) = \frac{P(o_1, \ldots, o_t|w_1, \ldots, w_t)P(w_1, \ldots, w_t)}{P(o_1, \ldots, o_t)}$$

# HMM Mathematical Model

- From the Markov property:

$$P(w_1, \dots, w_t) = \prod_{i=1}^{t} P(w_i | w_{i-1})$$

- Independent observations assumption:

$$P(o_1, \dots, o_t | w_1, \dots, w_t) = \prod_{i=1}^{t} P(o_i | w_i)$$



Some **assumption of independent observations** definitions are
1 "the occurrence of one event doesn't change the probability for another".
2 "sampling of one observation does not affect the choice of the second observation"
3 "Two events are independent if and only if $P(a \cap b) = P(a) * P(b)$
.

# HMM Mathematical Model

- Thus:
- $P(o_1, \dots, o_t | w_1, \dots, w_t) \propto \prod_{i=1}^{t} P(o_i|w_i) \prod_{i=1}^{t} P(w_i|w_{i-1})$

**HMM Parameters:**

- Transition probabilities $P(w_i|w_{i-1})$
- Emission probabilities $P(o_i|w_i)$
- Initial state probabilities $P(w_i)$

# HMM Parameters

- HMM is governed by the following parameters:

$$\lambda = \{A, B, \pi\}$$

  - State-transition probability matrix A
  - Emission/Observation/State Conditional Output probabilities $B$
  - Initial (prior) state probabilities $\pi$

- Determine the fixed number of states ($N$):

$$S = \{s_1, \dots, s_N\}$$

# HMM Parameters

- State-transition probability:

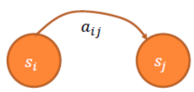
$$A = \begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1N} \\ a_{21} & a_{23} & \cdot & \cdot & \cdot & a_{2N} \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ a_{N1} & a_{N2} & \cdot & \cdot & \cdot & a_{NN} \end{bmatrix}$$

$\sum_{j=1}^{N} a_{ij} = 1$ (Each row/Outgoing arrows)

$a_{ij} = P(q_t = s_j \mid q_{t-1} = s_i), \quad 1 \le i,j \le N$

$a_{ij} \ge 0$

$a_{ij} \rightarrow \textit{Transisiton probability from state } s_i \textit{ to } s_j$



# HMM Parameters

- Emission probability distribution: A state will generate an observation (output), but a decision must be taken according on how to model the output, i.e., as discrete or continuous.
  - Discrete outputs are modeled using pmfs (*probability mass function*).
  - Continuous outputs are modeled using pdfs (*probability density function*).

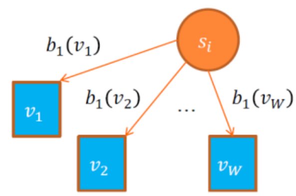
# HMM Parameters

- Discrete Emission Probabilities:

Observation Set: $V = \{v_1, \ldots, v_W\}$

$$b_i(v_k) = P(o_t = v_k | q_t = s_i), \qquad 1 \leq k \leq W$$

$$B = \begin{bmatrix} b_1(v_1) & b_1(v_2) & \cdot & \cdot & \cdot & b_1(v_W) \\ b_2(v_1) & b_2(v_2) & \cdot & \cdot & \cdot & b_2(v_W) \\ \vdots & \vdots & & \ddots & & \vdots \\ b_N(v_1) & b_N(v_2) & \cdot & \cdot & \cdot & b_N(v_W) \end{bmatrix}$$



# HMM Parameters

- Initial (prior) probabilities: these are the probabilities of starting the observation sequence in state $qi$.

$$\pi = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \cdot \\ \cdot \\ \cdot \\ \pi_N \end{bmatrix}$$

$$\pi_i = P(q_1 = s_i), \qquad 1 \le i \le N$$

$$\sum_{i=1}^{N} \pi_i = 1$$

# Three Basic Problems of HMM

- **Problem 1:** Probability Evaluation

  Given the observation sequence $\boldsymbol{O} = (\boldsymbol{o}_1 \boldsymbol{o}_2 \cdots \boldsymbol{o}_T)$, and a model $\lambda = (A, B, \pi)$, how do we efficiently computer $P(\boldsymbol{O}|\lambda)$, the probability of the observation sequence, given the model?

  - How do we compute the probability that the observed sequence was produced by the model?
  - Scoring how well a given model matches a given observation sequence.

# Three Basic Problems of HMM

- **Problem 2:** Optimal State Sequence

  Given the observation sequence $\boldsymbol{O} = (\boldsymbol{o}_1 \boldsymbol{o}_2 \cdots \boldsymbol{o}_T)$, and the model $\lambda$, how do we choose a corresponding state sequence $\boldsymbol{q} = (q_1 q_2 \cdots q_T)$ that is optimal in some sense (i.e., best explains the observations)

  - Attempt to uncover the hidden part of the model –that is, to find the “correct” state sequence.
  - For practical situations, we usually use an optimality criterion to solve this problem as best as possible.

# Three Basic Problems of HMM

- **Problem 3**: Parameter Estimation

  How do we adjust the model parameters $\lambda = (A, B, \pi)$ to maximize $P(\boldsymbol{O}|\lambda)$

  - Attempt to optimize the model parameters to best describe how a given observation sequence comes about.
  - The observation sequence used to adjust the model parameters is called a training sequence because it is used to “train” the HMM.

# HMM Example: Coins & Dice

$P(H|Red\ Coin) = 0.9$
$P(T|Red\ Coin) = 0.1$

$P(H|Green\ Coin) = 0.95$
$P(T|Green\ Coin) = 0.05$

$Outputs = \{1,2,3,4,5,6\}$


$Outputs$={1,1,1,1,1,1,1,2,3,4,5,6}

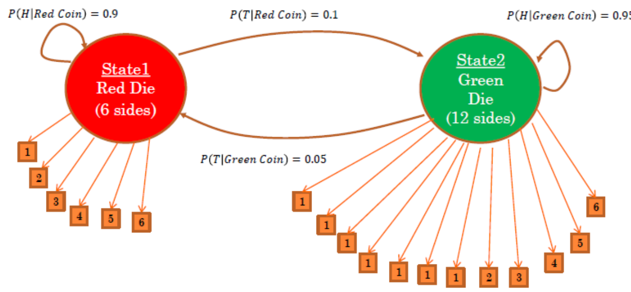
http://www.mathworks.com/help/stats/hidden-markov-models-hmm.html



The model creates a sequence of numbers from the set {1, 2, 3, 4, 5, 6} with the following rules:

- Begin by rolling the red dice and writing down the number that comes up, which is the emission.
- Toss the red coin and do one of the following:
  - If the result is heads, roll the red dice and write down the result.
  - If the result is tails, roll the green dice and write down the result.
- At each subsequent step, you flip the coin that has the same color as the dice you rolled in the previous step. If the coin comes up heads, roll the same dice as in the previous step. If the coin comes up tails, switch to the other dice.

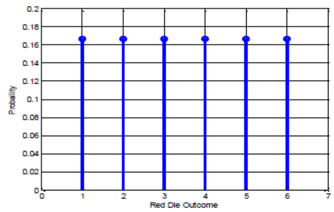
# HMM Example: Coins & Dice



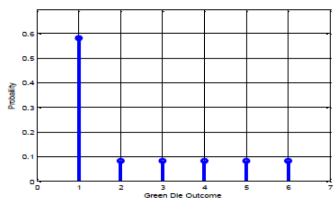

$$A = \begin{bmatrix} 0.9 & 0.1 \\ 0.05 & 0.95 \end{bmatrix} \qquad \pi = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

http://www.mathworks.com/help/stats/hidden-markov-models-hmm.html

# HMM Example: Coins & Dice



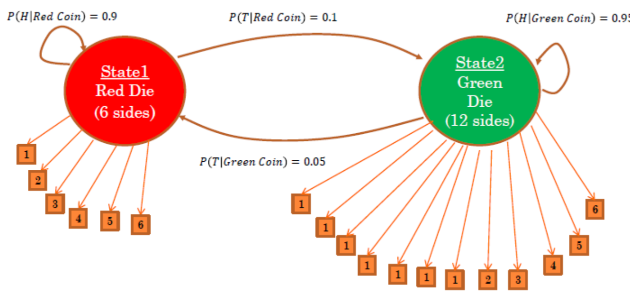

$$b_1(o_t) = \{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\}$$




$$b_2(o_t) = \{\frac{7}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}\}$$

http://www.mathworks.com/help/stats/hidden-markov-models-hmm.html

# HMM Example: Coins & Dice




$$A = \begin{bmatrix} 0.9 & 0.1 \\ 0.05 & 0.95 \end{bmatrix} \qquad \pi = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad B = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{7}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} \end{bmatrix}$$

http://www.mathworks.com/help/stats/hidden-markov-models-hmm.html

# Types of HMMs

- Ergodic
- Left-right
- Parallel path left-right

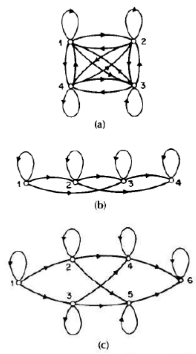



Fig. 7. Illustration of 3 distinct types of HMMs. (a) A 4-state ergodic model. (b) A 4-state left-right model. (c) A 6-state parallel path left-right model.

# Case Study

- Features
  - Field descriptor
  - Edge descriptor
  - Grass and sand
  - Player height

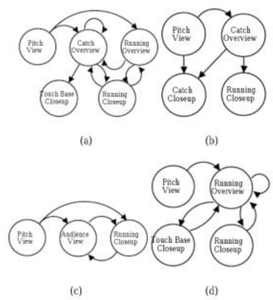



**Fig. 2.** (a) HMM model for nice hits (b) HMM model for nice catches (c) HMM model for home runs (d) HMM model for plays within the diamond

P. Chang, M. Han, and Y. Gong. "*Extract Highlights From Baseball Game Video With Hidden Markov Models*," In Proc. of ICIP, vol. 1, pp. 609-612, 2002.

# Summary

- Reasoning under uncertainty is an important area of AI
- It is not the case that statistical methods are the only way
- Logics can also cope with uncertainty in a qualitative way
- But statistical methods, and particularly Bayesian reasoning has become a cornerstone of modern AI (rightly or wrongly)
- A Markov Model is a stochastic model which models temporal or sequential data, i.e., data that are ordered.
- Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobservable (i.e. hidden) states.

# References


- S. J. Russell and P. Borvig, *Artificial Intelligence: A Modern Approach (4$^{th}$ Edition)*, Prentice Hall International, 2020.
  - Chapter 12. Quantifying Uncertainty
  - Chapter 13. Probabilistic Reasoning
  - Chapter 14. Probabilistic Reasoning over Time
- E. Fosler-Lusier, "*Markov Models and Hidden Markov Model: A Brief Tutorial*," International Computer Science Institute, 1998.
- L.R. Rabiner, "*A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*," Proceedings of the IEEE , vol.77, no.2, pp.257-286, Feb 1989
- John R. Deller, John, and John H. L. Hansen. "*Discrete-Time Processing of Speech Signals*". Prentice Hall, New Jersey, 1987.
- Barbara Resch (modified Erhard and Car Line Rank and Mathew Magimai-doss); "*Hidden Markov Models A Tutorial for the Course Computational Intelligence.*"
- Henry Stark and John W. Woods. "*Probability and Random Processes with Applications to Signal Processing (3rd Edition)*." Prentice Hall, 3 edition, August 2001.

# References


- S. J. Russell and P. Borvig, Artificial Intelligence: A Modern Approach (4th Edition), Prentice Hall International, 2020.

# eof