UNCERTAINTY

Unit 1

Syllabus – Unit 1

- **Uncertainty in AI**
- **Inference using full joint distributions**
- **Bayes Theorem**

-The semantics of Bayesian Networks
-Inference in Bayesian networks
- Decision Theory

- **Markov Decision Processes**.
- **Self-learning Topics:**

-Hidden Markov Model (HMM),
-Gaussian Mixture Model (GMM)

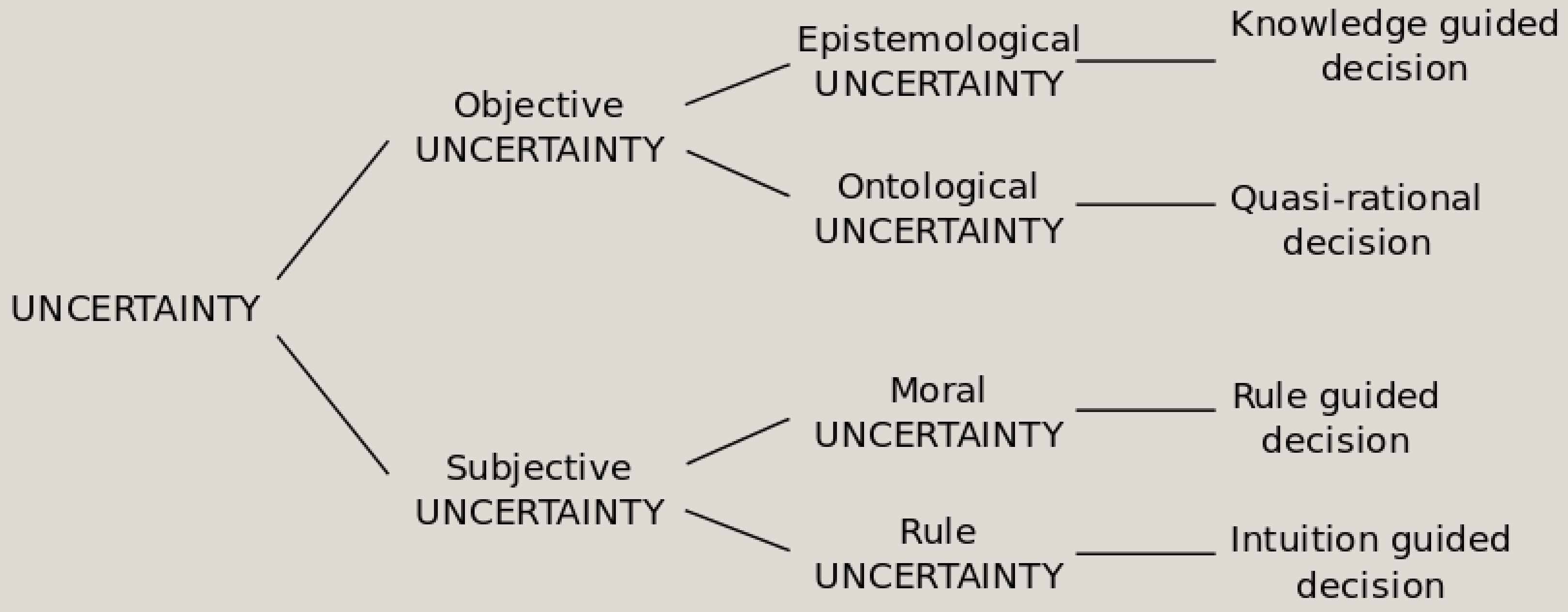
2 INTRODUCTION

- Uncertainty arises when we are not 100 percent sure about the outcome of the decisions.
- This mostly happens in those cases where the conditions are neither completely true nor completely false.

3 REASONS FOR UNCERTAINTY

- **Partially observable environment**
- **Dynamic environment**
- **Incomplete knowledge of the agent**
- **Inaccessible areas in the environment**

4 TAXONOMY OF UNCERTAINTY



5 METHODS TO HANDLE UNCERTAINTY

- Fuzzy Logic
- Probabilistic Reasoning
- Hidden Markov Models
- Neural Networks

6 PROBABILISTIC REASONING

- Probability is the calculus of **gambling**.
- Probabilistic reasoning is a way of knowledge representation where we apply the concept of probability to indicate the uncertainty in knowledge.
- Probability handles uncertainty that is the result of someone's laziness and ignorance.

7 A CLASSIC EXAMPLE

Can Tweety fly???

- Birds typically fly
- Tweety is a bird.
- **Tweety flies**

Birds typically fly Penguins are birds

Penguins typically do not fly Tweety is a Penguin.

- **Tweety does not fly.**

8 NEED OF PROBABILISTIC REASONING IN AI

1. When there are unpredictable outcomes.
2. When specifications or possibilities of predicates becomes too large to handle.
3. When an unknown error occurs during an experiment.

PROBABILITY

- Probability can be defined as a chance that an uncertain event will occur.
- The value of probability always remains between 0 and 1 that represent ideal uncertainties.
- Each possible world ω is associated with a numerical probability P(ω) such tha

$$0 \le P(\omega) \le 1$$

$$\Sigma_{\omega \in \Omega} P(\omega) = 1$$

- Example: If we are about to roll two (distinguishable) dice, there are 36 possible worlds to consider: (1,1), (1,2),..., (6,6)

- P(ω) =1/36

10 AXIOMS IN PROBABILITY

- $0 \leq P(A) \leq 1$, where P(A) is the probability of an event A.
- P(A) = 0, indicates total uncertainty in an event A.
- P(A) = 1, indicates total certainty in an event A.
- We can find the probability of an uncertain event by using the below formula.

$$\text{Probability of Occurrence} = \frac{\text{Number of desired outcomes}}{\text{Total number of outcomes}}$$

- $P(A')$ = probability of event A not happening.
- $P(A') + P(A) = 1$.

11 TERMINOLOGIES IN PROBABILITY

- Event
- Sample space
- Random variable
- Prior probability
- Posterior probability
- Conditional probability $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$

The conditional probability is not symmetrical : P(A | B) != P(B | A)

where, P(A ∩ B) = Joint Probability of A and B, P(B) = Marginal Probability of B and P(B) > 0

The joint probability is symmetrical : P(A, B) = P(B, A)

12 INFERENCE USING FULL JOINT DISTRIBUTIONS

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

- Probabilistic inference:The computation of posterior probabilities for query propositions given observed evidence.
- The full joint probability distribution specifies the probability of each complete assignment of values to random variables.
- **Marginalization:** to get the **marginal probability--** attained by adding the entries in the corresponding rows or columns
- For example, **P(cavity) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2**
- There are six atomic events for **(cavity ∨ toothache): 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28**
- **Variant of marginalization is called conditioning.**
- Computing a conditional probability $P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)} = \frac{0.108+0.012}{0.108+0.012+0.016+0.064} = \frac{0.12}{0.2} = 0.6$
- Similarly $P(\neg cavity \mid toothache) = \frac{0.016+0.064}{0.2} = 0.4$
- In both the cases, $\frac{1}{P(toothache)} = \frac{1}{0.2} = 5$ remains constant, no matter which value of cavity we calculate.
- It is a **normalization constant (α)** ensuring that the distribution P(cavity | toothache) adds up to 1.

EXAMPLE 1

In a class, there are 80% of the students who like English and 30% of the students who likes English and Mathematics, and then what is the percentage of students those who like English, also like mathematics?

E → English

M → Mathematics

$P(E) = 0.8$ $P(E \cap M) = 0.3$

To find :- $P(M|E) = \frac{P(E \cap M)}{P(E)} = \frac{0.3}{0.8} = 0.375$

$\underline{\underline{37.5\%}}$

EXAMPLE 2

The table below shows the occurrence of diabetes in 100 people. Let D and N be the events where a randomly selected person "has diabetes" and "not overweight". Then find P(D | N).

$$P(D|N) = \frac{P(D \cap N)}{P(N)} = \frac{5/100}{(5+45)/100} = \frac{5}{50} = 0.1$$

15 BAYES THEOREM

- Bayes' Theorem, named after 18th-century British mathematician Thomas Bayes.
- Bayes theorem plays a critical role in probabilistic learning and classification.
- Bayes' Theorem allows you to update the predicted probabilities of an event by incorporating new information.
- Uses *prior* probability of each category given no information about an item.
- It is often employed in finance in calculating or updating risk evaluation.
- The theorem is also called Bayes' Rule or Bayes' Law.

16 BASIC PROBABILITY FORMULAS

- Product rule
- Sum rule $$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$
- Bayes theorem $$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$
- Theorem of total probability, if event *Ai* is mutually exclusive and probability sum to one

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i)$$

17 BAYES THEOREM

- Given a hypothesis *h* and data *D* which bears on the hypothesis:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- *P(h)*: independent probability of *h*: *prior probability*
- *P(D)*: independent probability of *D*
- *P(D|h)*: conditional probability of *D* given h: *likelihood*
- *P(h|D)*: conditional probability of *h* given *D*: *posterior probability*

EXAMPLE 3

In Orange County, 51% of the adults are males. One adult is randomly selected for a survey involving credit card usage.

a. Find the prior probability that the selected person is a male.

b. It is later learned that the selected survey subject was smoking a cigar. Also, 9.5% of males smoke cigars, whereas 1.7% of females smoke cigars. Use this additional information to find the probability that the selected subject is a male.

$M \to$ Male $\quad \bar{M} \to$ Female (i.e. not male)

$C \to$ cigar smoker $\quad \bar{C} \to$ not a cigar smoker

a) $\boxed{P(M) = 0.51}$ $\quad \therefore P(\bar{M}) = 1 - P(M) = 1 - 0.51 = 0.49$

$P(C|M) = 0.095$

$P(C|\bar{M}) = 0.017$

b) $P(M|C) = \dfrac{P(C|M) \times P(M)}{P(C|M) \times P(M) + P(C|\bar{M})\, P(\bar{M})}$

$= \dfrac{0.095 \times 0.51}{0.095 \times 0.51 + 0.017 \times 0.49} = \dfrac{0.04845}{0.05678}$

$= \underline{\underline{0.8533}}$

$\boxed{P(M|C) = 0.8533}$

EXAMPLE 4

A doctor is called to see a sick child.The doctor has prior information that 90% of sick children in that neighborhood have the flu, while the other 10% are sick with measles. A well-known symptom of measles is a rash (the event of having which we denote R).

Assume that the probability of having a rash if one has measles is P(R | M) = 0.95. However, occasionally children with flu also develop rash, and the probability of having a rash if one has flu is P(R | F) = 0.08. Upon examining the child, the doctor finds a rash. What is the probability that the child has measles?

F → Flue

M → Measles

$P(F) = 0.90$

$P(M) = 0.10$

$P(R|M) = 0.95$

$P(R|F) = 0.08$

To find :- $P(M|R) = \dfrac{P(R|M) \cdot P(M)}{P(R|M) \cdot P(M) + P(R|F)\, P(F)}$

$= \dfrac{0.95 \times 0.10}{0.95 \times 0.10 + 0.08 \times 0.90} = \dfrac{0.095}{0.095 + 0.072}$

$= 0.5688$

INDEPENDENCE

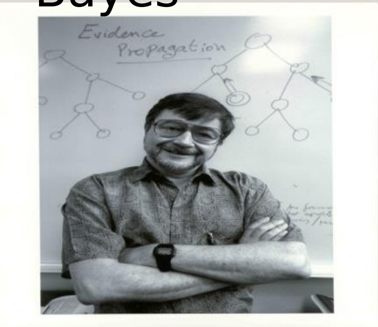
- Also called marginal independence / absolute independence.
- Reduce the amount of information necessary to specify the full joint distribution.
- Independence between variables X and Y can be written as:

$$\mathbf{P}(X|Y) = \mathbf{P}(X) \text{ or } \mathbf{P}(Y|X) = \mathbf{P}(Y) \text{ or } \mathbf{P}(X, Y) = \mathbf{P}(X)\mathbf{P}(Y)$$

BAYESIAN NETWORK MOTIVATION

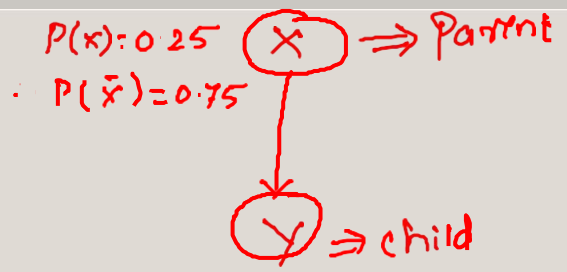
- We want a representation and reasoning system that is based on conditional independence
  - Compact yet expressive representation
  - Efficient reasoning procedures
- Bayesian Networks are such a representation
  - Named after Thomas Bayes
  - Term coined in 1985 by Judea Pearl
  - Their invention changed the focus on AI from logic to probability!

Thomas Bayes



Judea Pearl

22 BAYESIAN NETWORKS



- A Bayesian network specifies a joint distribution in a structured form
- Represent dependence/independence via a directed graph
  - Nodes = random variables
  - Edges = direct dependence
- Structure of the graph ⇔ Conditional independence relations
- Requires that graph is acyclic (no directed cycles)
- Two components to a Bayesian network
  - The graph structure (conditional independence assumptions)
  - The numerical probabilities (for each variable given its parents)

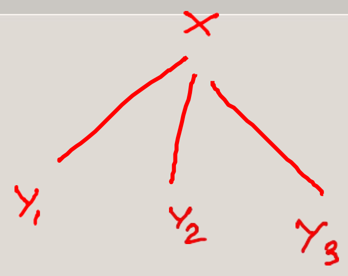
23 BAYESIAN NETWORKS

- General form:

$$P(X_1, X_2, \ldots . X_N) = \prod_i P(X_i \mid parents(X_i))$$

The full joint distribution

The graph-structured approximation

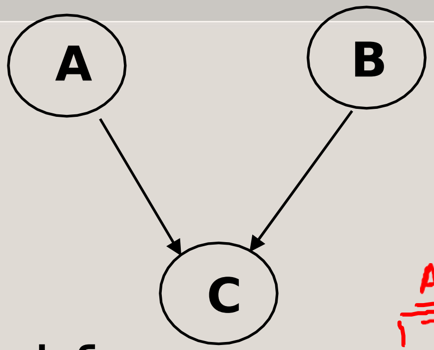


$$P(Y) = P(Y_1 | X)\, P(Y_2 | X)\, P(Y_3 | X)$$

24 EXAMPLE OF A SIMPLE BAYESIAN NETWORK

$$P(X_1, X_2, \ldots . X_N) = \prod_i P(X_i \mid parents(X_i))$$

$$\cancel{P(A, B, C)} = P(C \mid A, B)\, P(A)\, P(B) \longleftrightarrow$$



A	B	C	
T	T	T	F
T	F		
F	T		
F	F		

- Probability model has simple factored form
- Directed edges => direct dependence
- Absence of an edge => conditional independence
- Also known as belief networks, graphical models, causal networks

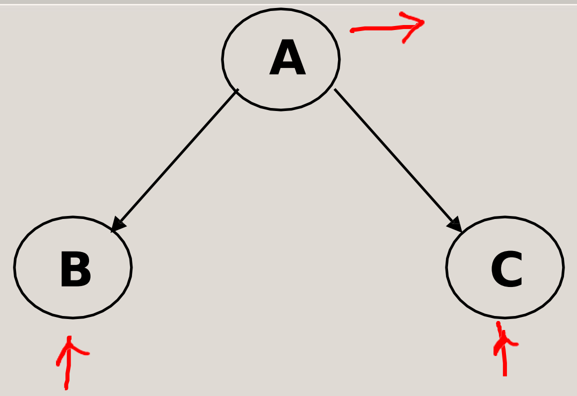
25 EXAMPLE OF 3-WAY BAYESIAN NETWORKS

- Conditionally independent effects:

$$p(A, B, C) = p(B|A)p(C|A)p(A)$$

- B and C are conditionally independent given A

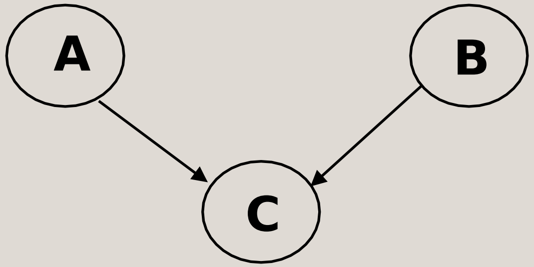
- e.g., A is a disease, and we model B and C as conditionally independent symptoms given A



26 EXAMPLE OF 3-WAY BAYESIAN NETWORKS

- Independent Clauses:

$$p(A, B, C) = p(C|A, B)p(A)p(B)$$

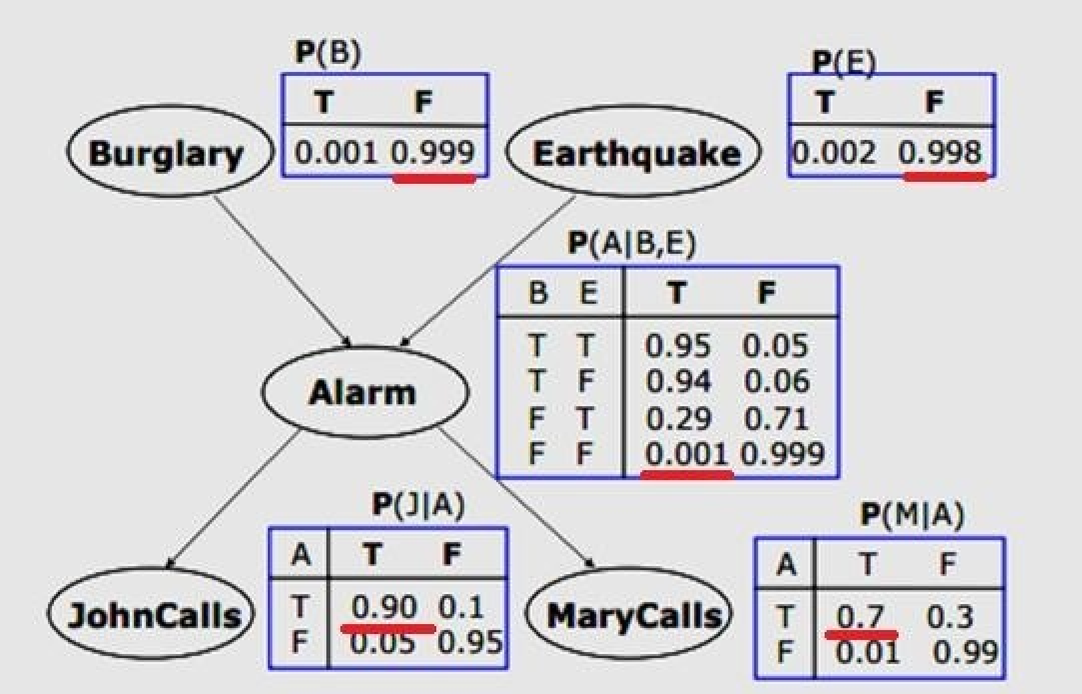


- “Explaining away” effect:
  - A and B are independent but become dependent once C is known!!

27 ALARM EXAMPLE

- Harry installed a new burglar alarm at his home to detect burglary. The alarm reliably responds at detecting a burglary but also responds for minor earthquakes. Harry has two neighbors John and Marry, who have taken a responsibility to inform Harry at work when they hear the alarm. John always calls Harry when he hears the alarm, but sometimes he got confused with the phone ringing and calls at that time too. On the other hand, Mary likes to listen to high music, so sometimes she misses to hear the alarm. Here we would like to compute the probability of Burglary Alarm.
- **Problem**: Calculate the probability that alarm has sounded, but there is neither a burglary, nor an earthquake occurred, and John and Mary both called the Harry.

SOLUTION



- **List of all events occurring in this network:**
- Burglary (B)
- Earthquake(E)
- Alarm(A)
- John Calls(J)
- Marry calls(M)
- We can write the events of problem statement in the form of probability:

$$P(M, J, A, \bar{B}, \bar{E})$$

$$= P(M|A) \times P(J|A) \times P(A|\bar{B} \cap \bar{E}) \times P(\bar{B}) \times P(\bar{E})$$

$$= 0.70 \times 0.90 \times 0.001 \times 0.999 \times 0.998$$

$$= 0.00068045$$

29 INFERENCE IN BAYESIAN BELIEF NETWORKS

- A Bayesian Network can be used to compute the probability distribution for any subset of network variables given the values or distributions for any subset of the remaining variables.
- Unfortunately, exact inference of probabilities in general for an arbitrary Bayesian Network is known to be NP-hard.

30 ADVANTAGES OF BAYESIAN BELIEF NETWORK

- Intuitive, graphical, and efficient
- Accounts for sources of uncertainty
- Allows for information updating
- Models multiple interdependencies
- Includes utility and decision nodes

31 DISADVANTAGES OF BAYESIAN BELIEF NETWORK

- Not ideally suited for computing small probabilities
- Computationally demanding for systems with a large number of random variables
- Exponential growth of computational effort with increased number of states

Decision Theory : Markov decision process

In mathematics, a **Markov decision process** (**MDP**) is a discrete-time stochastic control process. It provides a mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of a decision maker. MDPs are useful for studying optimization problems solved via dynamic programming. MDPs were known at least as early as the 1950s;[1] a core body of research on Markov decision processes resulted from Ronald Howard's 1960 book, ***Dynamic Programming and Markov Processes***.[2] They are used in many disciplines, including robotics, automatic control, economics and manufacturing. The name of MDPs comes from the Russian mathematician Andrey Markov as they are an extension of Markov chains.

At each time step, the process is in some state s, and the decision maker may choose any action a that is available in state s. The process responds at the next time step by randomly moving into a new state s', and giving the decision maker a corresponding reward $R_a(s, s')$.

The probability that the process moves into its new state s' is influenced by the chosen action. Specifically, it is given by the state transition function $P_a(s, s')$. Thus, the next state s' depends on the current state s and the decision maker's action a. But given s and a, it is conditionally independent of all previous states and actions; in other words, the state transitions of an MDP satisfy the Markov property.

Markov decision processes are an extension of Markov chains; the difference is the addition of actions (allowing choice) and rewards (giving motivation). Conversely, if only one action exists for each state (e.g. "wait") and all rewards are the same (e.g. "zero"), a Markov decision process reduces to a Markov chain.

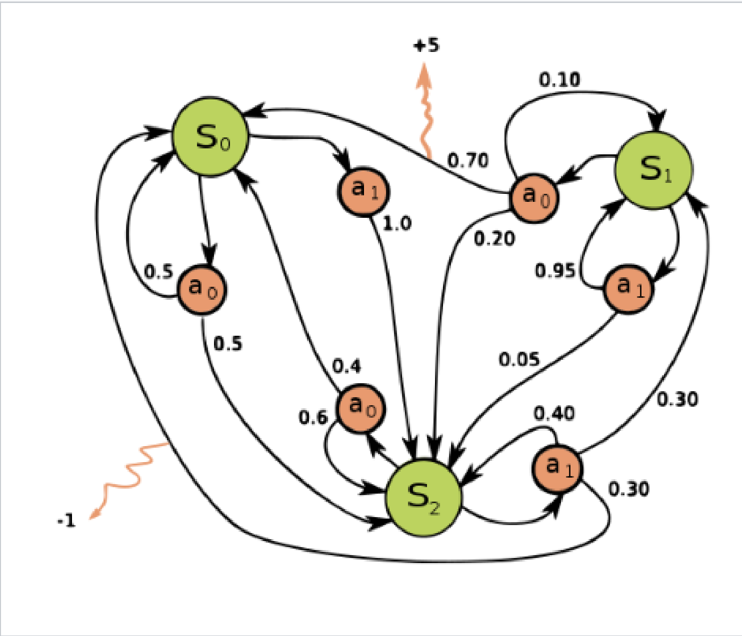
A Markov decision process (MDP)

A Markov decision process is a 4-tuple (S, A, P_a, R_a), where:

- S is a set of states called the *state space*,
- A is a set of actions called the *action space* (alternatively, A_s is the set of actions available from state s),
- $P_a(s, s') = \Pr(s_{t+1} = s' \mid s_t = s, a_t = a)$ is the probability that action a in state s at time t will lead to state s' at time $t+1$,
- $R_a(s, s')$ is the immediate reward (or expected immediate reward) received after transitioning from state s to state s', due to action a

The state and action spaces may be finite or infinite, for example the set of real numbers. Some processes with countably infinite state and action spaces can be reduced to ones with finite state and action spaces.[3]

A policy function π is a (potentially probabilistic) mapping from state space (S) to action space (A).



Example of a simple MDP with three states (green circles) and two actions (orange circles), with two rewards (orange arrows).

Optimization objective [edit]

The goal in a Markov decision process is to find a good "policy" for the decision maker: a function π that specifies the action $\pi(s)$ that the decision maker will choose when in state s. Once a Markov decision process is combined with a policy in this way, this fixes the action for each state and the resulting combination behaves like a Markov chain (since the action chosen in state s is completely determined by $\pi(s)$ and $\Pr(s_{t+1} = s' \mid s_t = s, a_t = a)$ reduces to $\Pr(s_{t+1} = s' \mid s_t = s)$, a Markov transition matrix).

The objective is to choose a policy π that will maximize some cumulative function of the random rewards, typically the expected discounted sum over a potentially infinite horizon:

$E\left[\sum_{t=0}^{\infty} \gamma^t R_{a_t}(s_t, s_{t+1})\right]$ (where we choose $a_t = \pi(s_t)$, i.e. actions given by the policy). And the expectation is taken over $s_{t+1} \sim P_{a_t}(s_t, s_{t+1})$

where γ is the discount factor satisfying $0 \leq \gamma \leq 1$, which is usually close to 1 (for example, $\gamma = 1/(1+r)$ for some discount rate r). A lower discount factor motivates the decision maker to favor taki actions early, rather than postpone them indefinitely.

A policy that maximizes the function above is called an *optimal policy* and is usually denoted π^*. A particular MDP may have multiple distinct optimal policies. Because of the Markov property, it can be show that the optimal policy is a function of the current state, as assumed above.

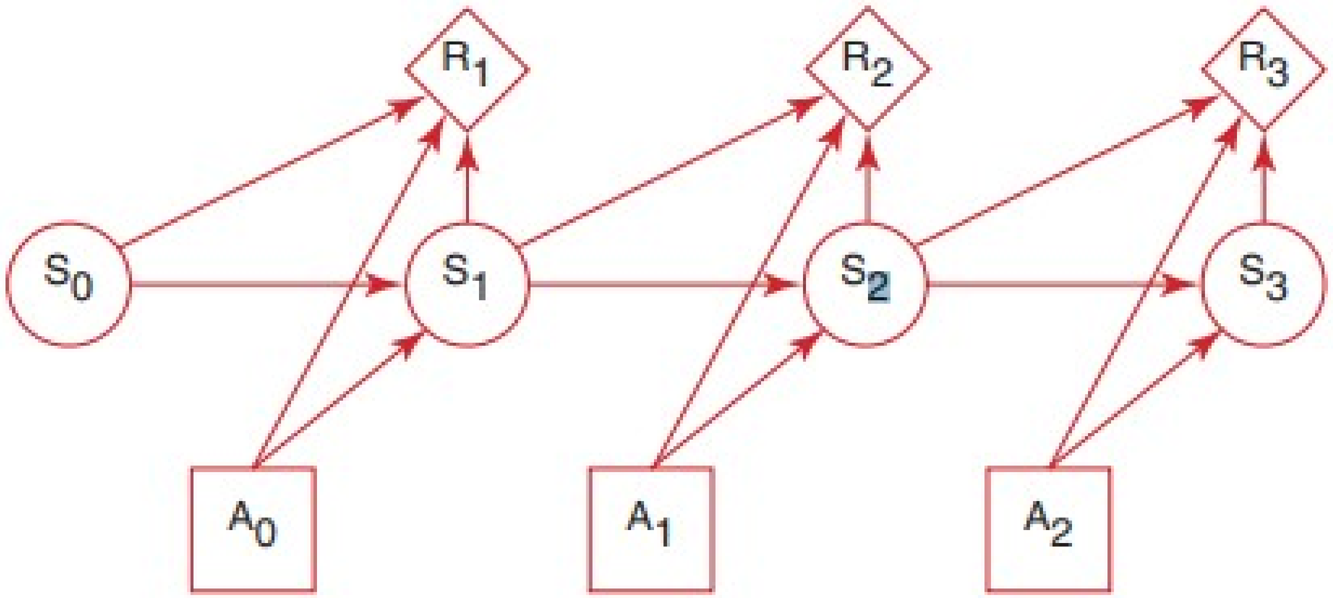
Markov Decision Processes

An MDP is defined by:

- set S of states.
- set A of actions.
- $P(S_{t+1}|S_t, A_t)$ specifies the dynamics.
- $R(S_t, A_t, S_{t+1})$ specifies the reward. The agent gets a reward at each time step (rather than just a final reward).
  - $R(s, a, s')$ is the reward received when the agent is in state s, does action a and ends up in state s'.

Decision Processes

- A Markov decision process augments a stationary Markov chain with actions and values:



Rewards and Values

Suppose the agent receives the sequence of rewards $r_1, r_2, r_3, r_4, \ldots$. What value should be assigned?

- total reward $V = \sum_{i=1}^{\infty} r_i$
- average reward $V = \lim_{n \to \infty} \frac{r_1 + \cdots + r_n}{n}$
- discounted reward $V = \sum_{i=1}^{\infty} \gamma^{i-1} r_i$
    - γ is the discount factor
    - $0 \leq \gamma \leq 1$

Policies

- A stationary policy is a function:

$$\pi : S \rightarrow A$$

  Given a state s, $\pi(s)$ specifies what action the agent who is following π will do.
- An optimal policy is one with maximum expected value
  - we'll focus on the case where value is defined as discounted reward.
- For an MDP with stationary dynamics and rewards with infinite or indefinite horizon, there is always an optimal stationary policy in this case.

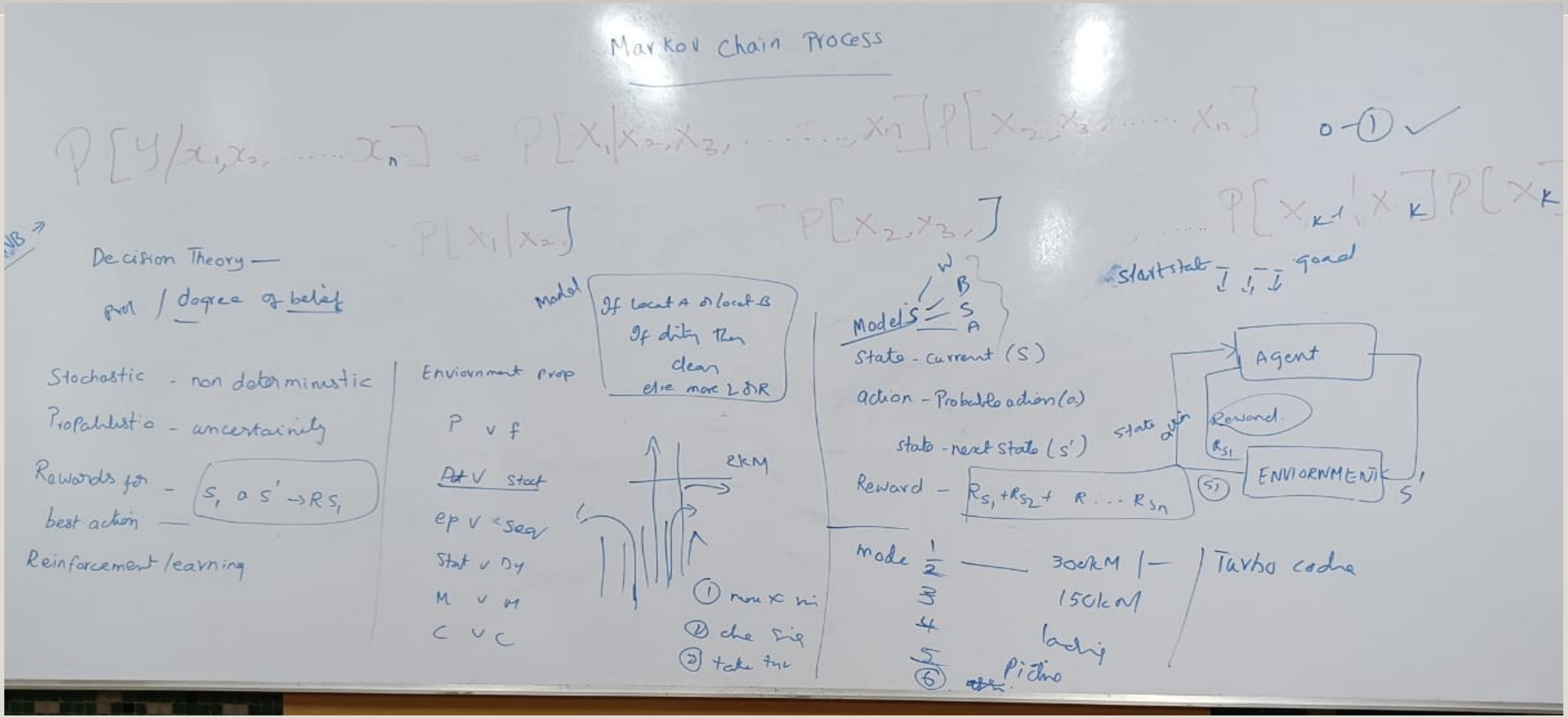
Pseudocode for Value Iteration

procedure value_iteration(P, r, θ)
inputs:
  P is state transition function specifying $P(s'|a, s)$
  r is a reward function $R(s, a, s')$
  θ a threshold $\theta > 0$
returns:
  $\pi[s]$ approximately optimal policy
  $V[s]$ value function
data structures:
  $V_k[s]$ a sequence of value functions
begin
  for $k = 1 : \infty$
    for each state s
      $V_k[s] = \max_a \sum_{s'} P(s'|a, s)(R(s, a, s') + \gamma V_{k-1}[s'])$
    if $\forall s \ |V_k(s) - V_{k-1}(s)| < \theta$
      for each state s
        $\pi(s) = \arg\max_a \sum_{s'} P(s'|a, s)(R(s, a, s') + \gamma V_{k-1}[s'])$
      return π, V_k
end

References

1. https://www.youtube.com/watch?v=oKko3ukFLVc&t=550s /* Bayes
2. https://www.javatpoint.com/bayesian-belief-network-in-artificial-intelligence#:~:text=We%20can%20define%20a%20Bayesian,decision%20network%2C%20or%20Bayesian%20model
3. https://en.wikipedia.org/wiki/Markov_decision_process
4. https://www.cs.ubc.ca/~kevinlb/teaching/cs322%20-%202006-7/Lectures/lect34.pdf

Markov Chain Process & Decision Thoery



Thank You