Decision Theory

9/4/25

Techniques to take decisions in the face of uncertainity:

- when faced with uncertainity, don't use ~~a.d~~ adhoc techniques.
- Decision theory provides well developed techniques
- But, what is uncertainity, & why do we need to reason with it?

Which route to take?

R_1 : S → F → B : 310

R_2 : S → R → P → B : 278

- Ideally would take shortest route.
- what if traffic is a factor now?
- suppose you have accurate info that
  - R1 has fast moving traffic while R_2 has slow moving traffic.
- A logical agent may reason
  - slow(R_2), fast(R_1)
  - slow(x) ⇒ Avoid(x)
  - Avoid(x) ^ fast(y) ⇒ select(y)
  - Agent selects R1.
- Do you think this is a good ~~re~~way to reason?
- why or why not?

Ex: 80% 4hrs 20% 10hrs 5.2hrs

70% 4.5hrs 30% 5hrs < 5hrs

80%

Uncertainity

- May not have categorical answers.
  - Maynot know exactly which route is slow.
  - 50% chance that one route is slow
  - Even if we know it is slow, how slow?
- Uncertainity changes the way an agent makes a decision.
- Probability theory used in helping agents reason with uncertainity.
  - Summarize uncertainity due to our ignorance or laziness.
  - May be very very costly to remove such uncertainity.
- Rational decision depends on
  - Relative importance of various goals
  - Likelihood or degree to which they will be achieved.

How to make decisions ?

Decision theory = Probability theory + (deals with chance)
utility theory (deals with outcomes).

- Fundamental idea's
  - The MEU (Maximum expected utility) Principle.

- Agent is rational if & only if it chooses the action that yields the highest expected utility averages over all possible outcomes of action.
- weigh the utility of each outcomes by the probability of that it occurs.

Utility theory – von Neumann (Game theory)

Let X be the set of outcomes. $\succ$ be the preference of a player over the set of outcomes.

Axioms:

- Completeness: every pair of outcomes is ranked.
- Transitivity: If $x_1 \succ x_2$ and $x_2 \succ x_3$ then $x_1 \succ x_3$
- Substitutability: If $x_1 \succ x_2$ then any lottery in which x_1 is substituted by x_2 is equally preferred.
- Decomposability: Two different lotteries assign same probability to each outcome, then player is indifferent b/w these two lotteries.
- Monotonicity: If $x_1 \succ x_2$ and $p > q$, then $[x_1 : p, x_2 : 1-p] \succ [x_1 : q, x_2 : 1-q]$
- Continuity: If $x_1 \succ x_2 \succ x_3$, $\exists p \Rightarrow x_2 \sim [x_1 : p, x_3 : 1-p]$

Von Neumann and Morgenstern:

Theorem: Given a set of outcomes X and a preferene relation on X that satisfies

above six axioms, there exists a utility function $u: X \longrightarrow [0,1]$ with the following properties:

$u(x_1) \geq u(x_2)$ iff $x_1 \succ x_2$

$U([x_1 : p_1; x_2 : p_2; \ldots; x_m : p_m]) = \sum_{j=1}^{m} p_j u(x_j)$

$\sum p_j = 1$

$= \sum_{j=1}^{m} p_j u(x_j)$

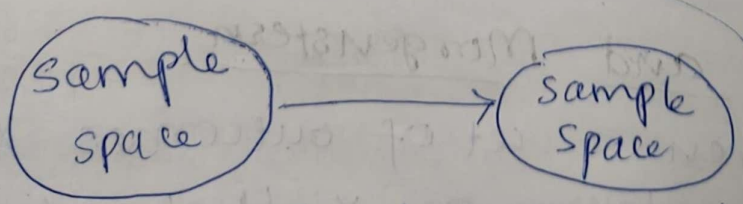
Basic Probability Notation

- Suppose we have prob. over all possible values (sample space) of a variable - denoted by $P(X)$.
- Called the prob. distribution for the random variable:
  - $P(\text{weather} = \text{sunny}) = 0.6$
  - $P(\text{weather} = \text{rain}) = 0.1$
  - $P(\text{weather} = \text{cloudy}) = 0.29$
  - $P(\text{weather} = \text{snow}) = 0.01$
  - $P(\text{weather}) = \langle 0.6, 0.1, 0.29, 0.01 \rangle$

$\sum P(X = x_i) = 1$

$x_i \in \{\text{sunny}, \text{rain}, \text{cloudy}, \text{sunny}\}$

Random variable:



→ This is not event space.

(Ω, f, P)

2^{Ω} - event space (discrete)

$2^{[0,1]}$ - continuous - Boral sigma algebra

Conditional Probability

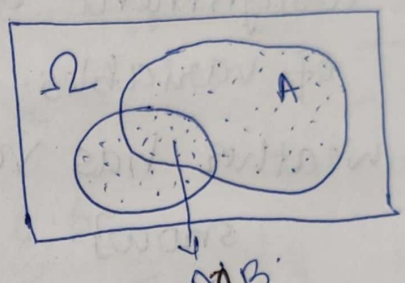
A, B

$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$ if $Pr(B) > 0$

$k_1 - (A \cap B)$

$k - B$

n - total sample space.

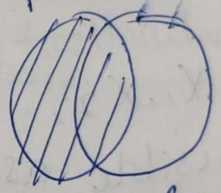


$\frac{k_1}{k} = \frac{k_1/n}{k/n} = \frac{Pr(A \cap B)}{Pr(B)}$

$A \cap B = A \wedge B$

Product Rule:

- $P(A|B) = P(A \wedge B) / P(B)$ if $P(B) > 0$ definition
- $P(A \wedge B) = P(A|B) * P(B)$
- $P(A \wedge B) = P(B|A) * P(A)$



Posterior or conditional probability

- $P(A|B)$ probability of A given all we know is B.
  - $P(x = \text{sunny} | \text{last 2 days were sunny})$.
- If we know B and also know C, then
  $P(A|B \wedge C)$.

Diagnosis domain: Introducing Cavity variable

- ~~Doctor~~ dia - photo.
- Not everytime.

Joint Probability Distribution

- Notation for distribution on multiple variables.
  - P(weather, cloudy) i.e., P(weather ∧ Cavity)
- Joint prob. distribution is a table.
  - Assign probabilities to all possible assignment of values for combinations of variables.
  - Weather has values {sunny, rain, cloudy, snow}.
  - Cavity has values {true, false}.
  - 4 * 2 table of probabilities called joint probability distribution.
- P($X_1, X_2, \ldots, X_n$) assigns probabilities to all possible assignment of values to variables $X_1, X_2, \ldots, X_n$.

Inference using full joint distribution

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012 ~~0.402~~	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

- 3 variables - Cavity, Toothache, Catch
- Dentist probe catches or not
- 2 * 2 * 2 table

$$Pr(X = x) = \sum_{(y,z)} Pr(X = x, Y = y, Z = z)$$

Marginalization

$P(\text{cavity}) = 0.108 + 0.012 + 0.072 + 0.008 =$
0.2 (called marginal probability of cavity).

$\Pr(\text{cavity} \mid \text{toothache}) = \frac{\Pr(C \wedge T)}{\Pr(T)} = \frac{0.12}{\Pr(T)}$.

$\Pr(T) = 0.108 + 0.012 + 0.016 + 0.064$
$= 0.108 + 0.080$

$0.12 + 0.08 = 0.20$

$= \frac{0.12}{0.2} = \frac{12}{20} = 0.6$

$\Pr(\text{cavity} \mid \text{toothache}) = 0.6$.

$\Pr(\neg\text{cavity} \mid \text{toothache}) = 0.4$ [1 − 0.6].

$= \frac{0.016 + 0.064}{\Pr(T)} = 0.4$.

$P(\text{cavity} \mid \text{toothache}) + P(\neg\text{cavity} \mid \text{toothache}) = 1$

Marginalization

- Denominator can be viewed as a normalization constant $1/P(\text{toothache})$.
- $P(\text{Cavity} \mid \text{Toothache}) = \alpha\, P(\text{cavity}, \text{toothache}) =$
  $\alpha [P(\text{cavity}, \text{toothache}, \text{catch}) +$
  $P(\text{cavity}, \text{toothache}, \neg\text{catch})] =$
  $0.108 + 0.012 = \alpha\, 0.12$.
- $\alpha\, P(\neg\text{cavity}, \text{toothache}) = \alpha\, [0.016 + 0.064] =$
  $= \alpha\, 0.08$
- Normalizing $\alpha \langle 0.12, 0.08 \rangle$ gives $\langle 0.6, 0.4 \rangle$

- Issue: For a domain with n Boolean variables, i/p table size of $O(2^n)$.
- Full joint distribution in tabular form not a practical tool for building reasoning systems.

Independence:

- Helps in reducing size of domain representation and complexity of inference problem.
- Let's add weather into the example: Sunny, rain, cloudy, snow → 2*2*2*4 entries in table.
- P(toothache, catch, cavity, cloudy) = P(cloudy | toothache, catch, cavity). P(toothache, catch, cavity).
- P(cloudy | toothache, catch, cavity) = P(cloudy)
- 32 element table becomes 8 element + 4 element table.
- $Pr(A|B) = \frac{Pr(A)\,Pr(B)}{Pr(B)} = Pr(A)$.

Bayes' Rule

- Given that
  - $P(A \wedge B) = P(A|B) * P(B)$
  - $P(A \wedge B) = P(B|A) * P(A)$

- $P(B|A) = P(A|B) \cdot P(B) / P(A)$.
- We can determine $P(B|A)$ given $P(A|B)$, $P(B)$ and $P(A)$.
- $P(\text{effect}|\text{cause})$ may be casual knowledge, $P(\text{cause}|\text{effect})$ diagnostic knowledge.
- Often have conditional probabilities on causal relationships $\longrightarrow P(\text{symptoms}|\text{disease})$
- Need to infer diagnostic knowledge many times.

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$$

$$Pr(B|A) = \frac{Pr(A \cap B)}{Pr(A)}$$

$$Pr(B|A) = \frac{Pr(A|B) \cdot Pr(B)}{Pr(A)}$$

$Pr(B|A)$ ↙ diagnostic

12|4|25

Probability theory

$$Pr(B|A) = \frac{Pr(A|B) \cdot Pr(B)}{Pr(A)}$$

$Pr(B|A)$ ↙ Posterior / Diagnostic knowledge

$Pr(A|B)$ ↙ Likelihood / Casual knowledge.

$Pr(B)$ → Prior

$Pr(A)$ → evidence

Ex:

$$Pr(I|F) = \frac{Pr(F|I) \cdot Pr(I)}{Pr(F)}$$

Ex: S: Proposition that patient has stiff neck
M: Proposition that patient has meningitis
Meningitis causes stiff-neck, 70% of the time

Given,

$P(S|M) = 0.7$

$P(M) = 1/50,000$

$P(S) = 0.01$

$$P(M|S) = \frac{P(S|M) \cdot P(M)}{P(S)} = \frac{0.7 \times \frac{1}{50,000}}{0.01}$$

$$= \frac{7 \times 10^{-1} \times \frac{1}{5} \times 10^{-4}}{10^{-2}} = 7 \times 10^{-1} \times 0.2 \times 10^{-2}$$

$$= 1.4 \times 10^{-3} = 0.0014$$

↓ because Pr. of meningitis is low

→ Why not estimate $P(M|S)$ the same way as we do $P(S|M)$? (low)

↳ Why do we need Baye's rule?

- $P(S|M)$ is causal knowledge, doesnot change
  - It is "model based"
  - It reflects the way meningitis works
- $P(M|S)$ is diagnostic, tells us likelihood of M given symptom S.
  - Diagnostic knowledge may change with circumstance, thus helpful to derive it.
  - If there is an epidemic, probability of meningitis goes up.

rather than trying to estimate P(M|S) which is very hard, we can compute it using the other terms.

- However, P(S|M) is unaffected by epidemic

P(E|C) ⟶ casual lang.

↓ because of this, this is effect(E)

Computing the denominator : P(S)

We wish to avoid computing the deno. in the Bayes rule

- May be hard to obtain
- In our example, the deno. is P(S)
  - Techniques to compute {or avoid computing P(S)}.

Our first approach is to compute relative likelihoods

If M(meningitis) and W(Whiplash) are two possible explanations:

- $P(M|S) = P(S|M) * P(M) / P(S)$
- $P(W|S) = P(S|W) * P(W) / P(S)$
- $\boxed{P(M|S) / P(W|S) = \frac{P(S|M) * P(M)}{P(S|W) * P(W)}}$

Another approach; Approach #2

- $P(M|S) = P(S|M) * P(M) / P(S)$
- $P(\text{Not}(M)|S) = P(S|\text{Not}(M)) * P(\text{Not}(M)) / P(S)$

- P(M|S) + P(Not(M)|S) = 1 [these 2 quantities must sum to 1].
- [P(S|M) * P(M) / P(S)] + [P(S|Not(M)) * P(Not(M)) / P(S)] = 1
- P(S|M) * P(M) + P(S|Not(M)) * P(Not(M)) = P(S)
- Calculate P(S) in this way?

Bayes $$P(B|A_i) = \frac{P(A_i|B)\,P(B)}{\sum_{i=1}^{k} P(A_i|B)\cdot P(A_i)}$$

$\cup A_i = \Omega$, $A_i \cap A_j = \phi$.

- P(M|S) + P(W|S) + P(D|S) + P(ND|S) = 1

Simple Example:

- Suppose 2 identical urns
- URN1 colored red from inside, has 1/3 black balls, 2/3 red balls
- URN2 colored black from inside has 1/3 red balls, 2/3 black balls.
- We select one URN at random; can't tell how it is colored inside.
- What is the probability that URN is colored inside? 0.5
- What if we were to select a ball at random from URN, and it is red? Does that change the probability?

$P(\text{Red-urn} \mid \text{Red-ball}) = P(\text{Red-ball} \mid \text{Red-urn}) * \frac{P(\text{Red-urn})}{P(\text{red-ball})}$

$= \frac{\frac{2}{3} \times 0.5}{P(\text{red-ball})}$

→ How to calculate P(red-ball)?

$P(\text{Black-urn} \mid \text{Red-ball}) = P(\text{Red-ball} \mid \text{Black-urn}) * \frac{P(\text{Black urn})}{P(\text{Red-ball})}$.

$= \frac{\frac{1}{3} \times 0.5}{P(\text{red-ball})}$.

Thus by our approach#2 : $\frac{2}{3} \times 0.5 / P(\text{red-ball}) + \frac{1}{3} \times 0.5 / P(\text{red-ball}) = 1$

Thus, $P(\text{red-ball}) = 0.5$

$\therefore P(\text{Red-urn} \mid \text{red-ball}) = 2/3$

More General forms of Bayes' rule:

- $P(Y|X) = P(X|Y) * P(Y) / P(X)$
- Bayes' rule for multi-valued variables
- Generalize to some background evidence e
- $P(Y|X, e) = P(X|Y, e) * P(Y|e) / P(X|e)$

Conditional Independence

- Toothache & Catch are independent given the presence of Cavity.
- Each is directly caused by cavity. toothache is caused by cavity

and a dentist catches using probe.

- A toothache cannot be caught by a probe nor a probe results in a toothache
- Conditional independence captured as:

$P(\text{toothache} \wedge \text{catch} \mid \text{cavity}) = P(\text{toothache} \mid \text{catch}, \text{cavity}) * P(\text{catch} \mid \text{cavity})$.

- $P(\text{toothache} \mid \text{catch}, \text{cavity}) = P(A \mid B, C)$
  - If A & C are conditionally independent given B
  - Then, probability of A is not dependent on C.
  - i.e., $P(A \mid B \wedge C) = P(A \mid B)$
- Therefore, $P(\text{toothache}, \text{catch} \mid \text{cavity}) = P(\text{toothache} \mid \text{cavity}) \cdot P(\text{catch} \mid \text{cavity})$

A = toothache, B = Cavity, C = catch.

Combining Evidence

- S : Proposition that patient has stiff neck
- H : Proposition that patient has severe headache.
- M : Proposition that patient has meningitis
- Meningitis causes stiff-neck, 50% of the time.
- Meningitis causes head ache, 70% of the time
- Probability of meningitis should go up if both symptoms reported - how to

combine such

- $P(A|B \wedge C) = P(A \wedge B \wedge C) / P(B \wedge C)$
- Numerator:
  - $P(M \wedge S \wedge H) = P(S|M \wedge H) * P(M \wedge H)$.
    $= P(S|M) * P(M \wedge H)$ → Why?
    $= P(S|M) * P(H|M) * P(M)$.

Going back to our example:

$P(M|S \wedge H) = P(S|M) * P(H|M) * P(M) / P(S \wedge H)$.

S & H are conditionally independent

Romania example:

- If plan1 and plan2 are the two plans
  - Plan 1 uses route 1.
    - $P(\text{home-early} | \text{plan 1}) = 0.8$.
      $P(\text{stuck 1} | \text{plan 1}) = 0.2$
    - Route 1 will be quick if flowing, but stuck for 1 hour if slow.
      - $V(\text{home-early}) = 100$.
        $V(\text{stuck 1}) = -1000$
      - Assigned numerical values to outcomes
  - Plan 2 uses route 2.
    - $P(\text{home-somewhat-early} | \text{plan 2}) = 0.7$,
      $P(\text{stuck 2} | \text{plan 2}) = 0.3$
    - Route 2 will be somewhat quick if flowing, but not bad even if slow.
      $V(\text{home-somewhat-early}) = 50$, $V(\text{stuck 2}) = -10$

Appln of MEU principle

- $EU(\text{Plan 1}) = P(\text{home early} \mid \text{plan 1}) * U(\text{home early}) + P(\text{stuck 1} \mid \text{plan 1}) * U(\text{stuck 1})$

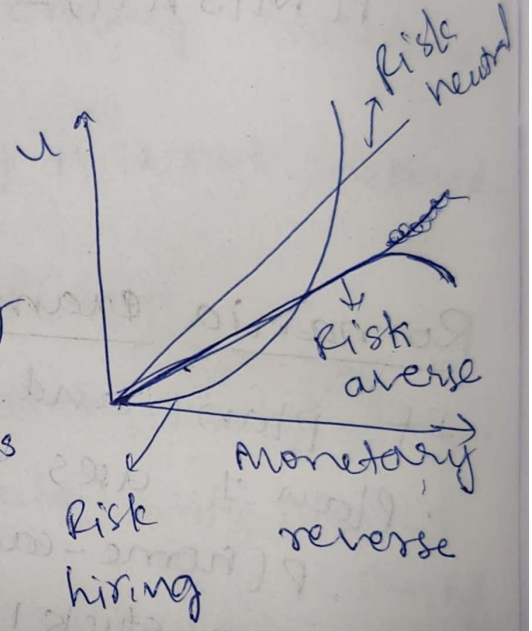
$= 0.8 \times 100 + 0.2 * (-1000) = -120$

- $EU(\text{plan 2}) = P(\text{home - somewhat - early} \mid \text{plan 2}) * U(\text{home - somewhat - early}) + P(\text{stuck 2} \mid \text{plan 2}) * U(\text{stuck 2})$

$= 0.7 \times 50 + 0.3 \times (-10)$

$= 32$

Risk Aversion

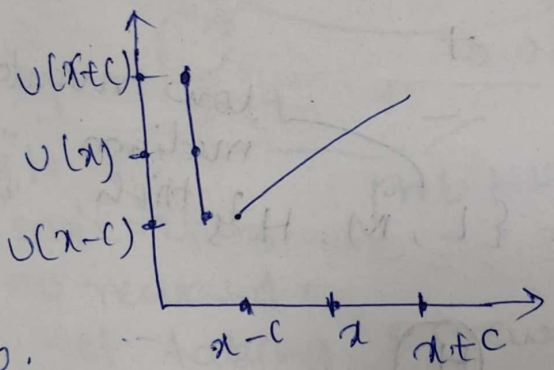


- We are risk averse
- Our utility f^n. are money as follows:
  - Our first million means alot & $U(\$1M) = 10$
  - Second million not so much $U(\$2M) = 15$ (not 20)
  - Third million even less so $U(\$3M) = 18$ (not 30)
  - ...
- Additional money is not buying us as much utility.
- If we plot amount of money on the x-axis and utility on the y-axis, we get a concave curve

More Risk Aversion

- Key: Slope of utility f^n is continuously decreasing.
- We will refuse to play a monetarily fair bet:

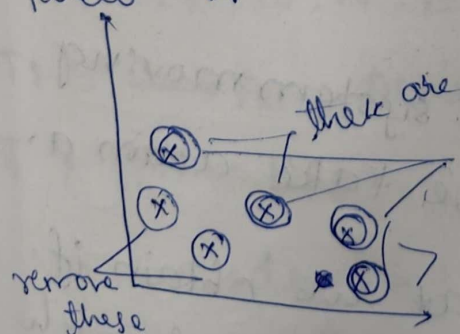
. Suppose we start with x dollars.
. We are offered a game:
  . 0.5 chance to win 1000 dollars (c=1000)
  . 0.5 chance to lose 1000 dollars (c=1000)
  . Expected monetary gain or loss is zero (hence monetarily fair).
. should be neutral to it, but seems we are not! why?
  . $U(x+c) - U(x) < U(x) - U(x-c)$
  . $U(x+c) + U(x-c) < 2U(x)$
  . $[U(x+c) + U(x-c)/2] < U(x)$
  . EU (Playing the game) < EU (not playing the game).



MEU:

16/4/25

→ Rule out dominated values.
→ The one which maximizes the weighted sum.
→ Pareto efficient — Eg: two features multiple options



these 3 are pareto efficient, we can select any one.
no one is dominating the others.

Utility theory —
- Risk averse
- Risk Neutral
- Risk seeking / loving

Pareto efficient:

Suppose we are dealing with N-dimensional utilities. Option x is pareto efficient if

$\nexists\, y$ s.t $u_i(y) > u_i(x)$ $i = 1, \ldots, N$

with strick inequality for atleast one i.

→ Pareto efficient is one method to select.

→ $f(u_1, \ldots, u_N)$ - mapping. Need not be linear. (one method of selection).

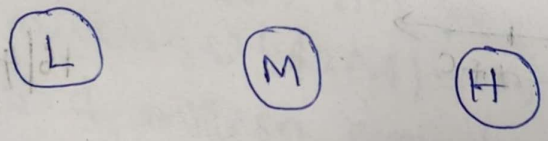
Actions:

Ex:

	Production	Marketing
a_0	+	+
a_1	+	−
a_2	−	+
a_3	−	−

Markov Decision process:

Let states be, $S = \{L, M, H\}$ — low inventory, medium, high



P : $S \times S \times A \longrightarrow [0, 1]$
(current, future)

R : $S \times S \times A \longrightarrow \mathbb{R}$
(current, destination, future state)

$P_a(s, s')$: probability of a system moving from state s to s', if we take action a.

$R_a(s, s')$: The reward that we obtain if we take an action a in state s & reach state s'.

$\langle S, A, R, P \rangle$ – Markov decision process (MDP)

- Depends on the previous step only, not on all previous state.
- Pa – markovian process
- Ra – not markovian, just showing a change in system, it is not a process

Example: Photo

N, E, S, W

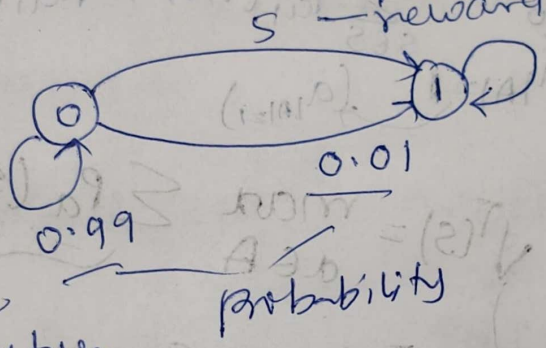
(1,3), (1,2), (1,1)

(2,3), (2,1)

(3,3), (3,2), (3,1)

Will MEU work here?

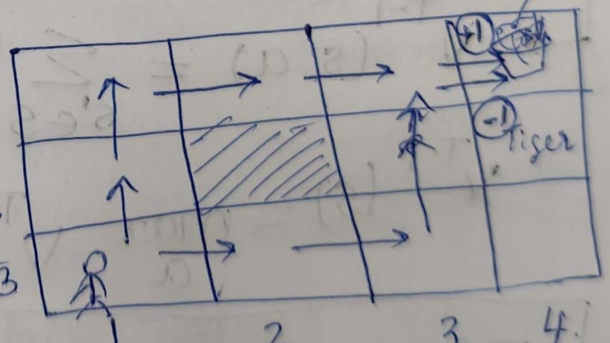
Ex:



- but after going to ①, we can't return & no reward.

So, not taking S (reward) & staying in '0' is better.

We can assure MEU is the best.



Rd: N, N, E, E, E

E, E, N, N, E

$\Pi : S \longrightarrow A$ Deterministic

$\Pi : S \longrightarrow D(A)$ Randomized.

- No randomized policy will give better than optimal deterministic function.

T rounds (finite horizon)

infinite horizon

~~$\Pi_0, \Pi_1, \Pi_2, \ldots$~~

$\Pi^0, \Pi^1, \Pi^2, \ldots$

$\Pi^0 = \Pi^1 = \Pi^2 \longrightarrow$ (stationary policy).

$|A|^{|S|}$ are possible policies.

$\Pi_0 = (0, a_0)(1, a_0)$, $\Pi_2 = (0, a_1)(1, a_0)$

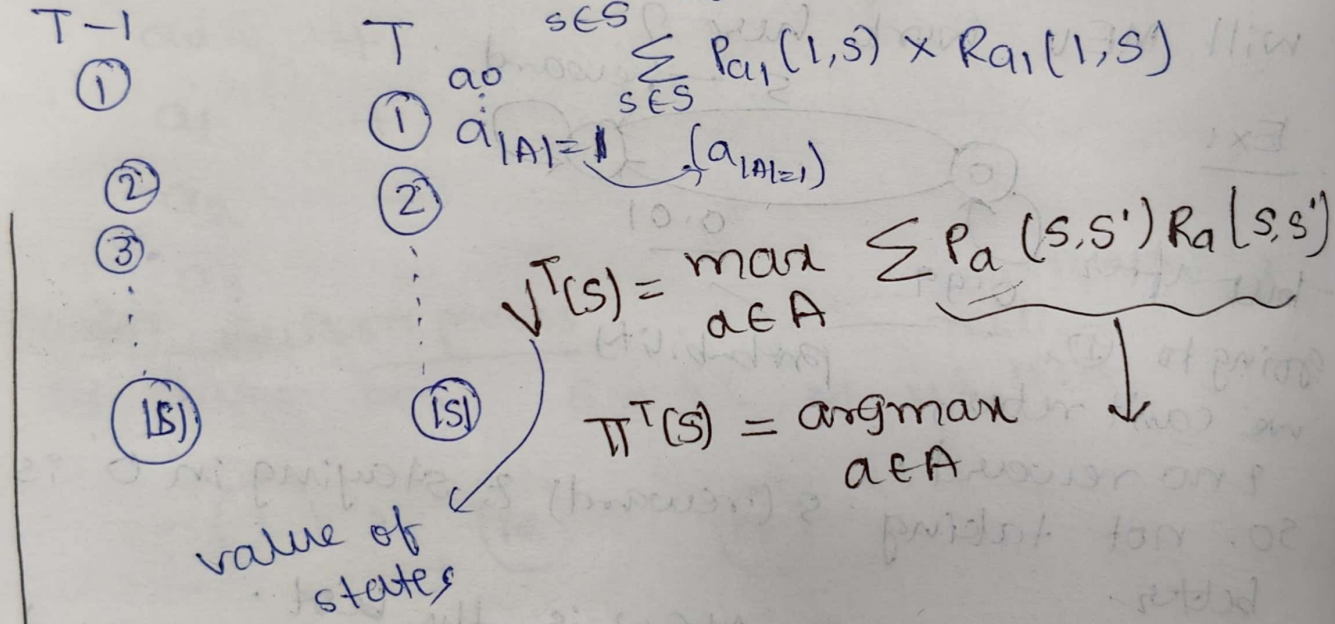
$\Pi_1 = (0, a_1)(1, a_1)$ $\Pi_3 = (0, a_0)(1, a_0)$

Finite horizon :

Each action gives some rewards (Expected reward)

$\sum_{s \in S} P_{a_0}(1, S) \times R_{a_0}(1, S)$

$\sum_{s \in S} P_{a_1}(1, S) \times R_{a_1}(1, S)$



$$V^{T-1}(s,a) = \sum_{s' \in S} P_a(s,s' [V^T(s') + R_a[s,s']]$$

$$V^{T-1}(S) = \max_a V^{T-1}(s,a)$$

$$V^{T-k}(S) = \max_a \left[\sum_{s' \in S} P_a(s,s') \{ V^{T-k+1}(s') + R_a(s,s') \} \right]$$

→ T is given.

① ①

a_0, a_1 $\quad P_{a_0}(0,1) = 0.6 = P_{a_1}(1,0)$

$\pi(0) = a_0$ $\quad P_{a_0}(0,0) = 0.4 = P_{a_1}(1,1)$

$\pi(1) = a_1$

$R_{a_0}(0,1) = R_{a_0}(0,0)$

$R_{a_1}(1,0) = R_{a_1}(1,1) = 1$

let $V(0) = 2$, $V(1) = 3$.

and consider $\pi(0) = a_0$, $\pi(1) = a_1$

these values depend upon policy.

$$V^{\pi}(0) = 0.4\,[1 + \underset{\downarrow}{2}] + 0.6\,[1 + \underset{\downarrow}{3}]$$

($V(0)$, $V(1)$)

$$V^{\pi}(0) = 3.6$$

$$V^{\pi}(1) = 0.4\,[1 + \underset{\downarrow}{3}] + 0.6\,[1 + \underset{\downarrow}{2}] = 4.36$$

($V(1)$, $V(0)$)

$R_{a_1}(1,0)$ $\qquad$ $R_{a_1}(1,1)$

→ 5, 5, 5, 5, 0, 5, 5, 5, 5, 0, . . . }

4, 4, 4, 4, 4, 4 . . . . } are both equal or the rewards diff?

→ 0, 0, 0, 0, 0, 5, 5, 5, 5, 0 . . .

4, 4, 4, 4, 4, 4, 4 . . . . } equal?

$\gamma = 0$ (ignoring future value)

$\gamma \rightarrow$ discount for the values.

generally b/w [0 & 1).

$$V^{\pi}(0) = 0.4\,[1 + \gamma 2] + 0.6\,[1 + \gamma 3]$$

↳ discount factor

$$V(s) = \max_a \sum_{s' \in S} P_a(s,s') \{ R_a(s,s') + \gamma V(s') \}$$

$\gamma < 1$:

as $t \to \infty$ $\| V^{t+1}(s) - V^t(s) \|_2 \to 0$

- Algo. converges.
- Value iteration.

Monte-carlo method
Q-learning
SARSA

Double Q
DQN
DDQN
DDPG

Markov Decision Process (MDP) 19/4/25

(S, A, R, P)

Finite Horizon MDP

- Maximum expected utility.
- If we are in last, it doesn't matter what is the last state.

$$V^{T-1}(s) = \sum_{s'} \{ R_a(s,s') + \gamma \cdot V^T(s') \} P_a(s,s')$$

$\gamma < 1$

$\gamma = 0$, don't care about future rewards

$$V^{T+1}(s) = \max_a \sum_{s'} P_a(s,s') \{ R_1(s,s') + V^t(s') \}$$

this converges.

Infinite Horizon MDP

→ we have only one state $|S|=1$. What policy should we use?

→ Maximum expected utility is used → optimal policy to use.

$$\pi^*(s) = \underset{a}{\operatorname{argmax}}\ R_a(s,s')$$

If we know that

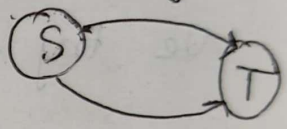
$R_a = 1$ w.p M_a

$= 0$ otherwise

$a^* = \operatorname{argmax} M_a$.

→ If we don't know M_a?

We have 2 states S, T. There are some delays in the path. Which one would you choose?



- Multi-armed Bandits (MAB).
- Check all possible routes, find the one which has ~~its~~ minimum delay. This is a greedy approach.

Greedy Approach (ε):

for $i = 1 \rightarrow mk$ } Exploration
  use action i
  observe R_i
  update $\hat{u}_i$

for $i = mk+1 \rightarrow T$ } Exploitation
  use action $i^* = \operatorname{argmax} \hat{u}_i$
  update $\hat{u}_{i^*}$

Issues:

→ A new smaller reward can dominate in first trial.

$\mu_1 = 0.9$, $\mu_2 = 0.8$, $m = 1$

$\hat{\mu}_1 = 0$, $\hat{\mu}_2 = 1$

$m = T/k$

- Using Round Robin

One every trial, $\left[\frac{0.1 \times T}{2}\right]$

Exploitation: → ε_t Greedy Algorithm

w.p $(1-\varepsilon)$ use greedy strategy

w.p ε select an action randomly (w.p $1/k$)

↓ learning rate

→ How to set ε?

We try to find $\varepsilon_t = O(1/t)$

as time increase, ε_t decreases.

→ How do we compare diff. algorithms?

Regret.

Expected loss in reward

$$\sum_{t=1}^{T} (\mu_a^* - \mu_{at})$$

→ ~~$O(\log T)$~~ . $\Omega(\log T)$ (lower bound)

→ No algorithm can do better than this.

Upper Confidence Bound (UCB) - Confidant also than greedy

→ $O(\log T \cdot \sqrt{\log T})$.

for t=

use (pull arm) action t

update $\hat{\mu}_t^*$

for $t = k+1 \rightarrow T$

pull ~~$t = a$~~ $a^t = \text{argmax}\ \hat{\mu}_i^t + \sqrt{\frac{2\ln t}{N_{i,t}}}$

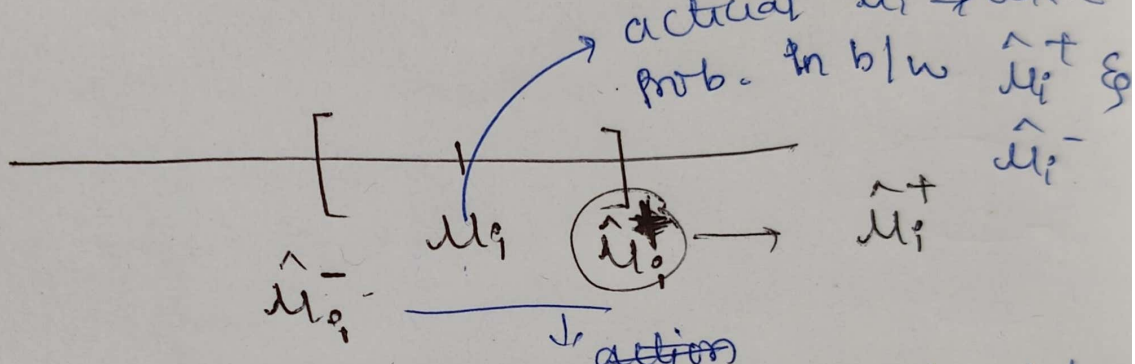
UCB

$\hat{\mu}_i^t$ ↓ imperical mean

~~$N(i,t)$~~

→ $N_{i,t}$ = Number of times action/arm is used till time t.

$\hat{\mu}_i^t$ ↓



- when $a^t \neq a^*$, we will add reward.
- $O(\log T)$. → since need to pull T times. no better also. to pull $< T$ times.
- $N_{i,t}$ is low, the interval will be high.

↓ goes $(\log T)$ times.