MDP's	POMDP's	Dec ~~POMDP's~~ POMDP's (Decentralized)
Similarity: Structure ie Agent & environment ~~Goals~~		
Differences: 1 Agent: Environment is known * Goals are generalized reward f^n's. * Actions are stochastic. * Perfect perception ⇓ Atomic Model	1 Agent but environment is partially known ie partial observability	Other agents + POMDP's

Idea from Self

A side note from an idea in Theory of Computation

* We know that given a finite memory we could only perform if the problem could be expressed in finite states
* Therefore if we could express the problem in a finite state space we could solve it ; even if it takes non-polynomial time.

→ Exponential ⇒ finite horizon cases

→ Undecidable ⇒ in infinite horizon cases

Why Dec. MDP's exists? ⇒ If multiple agents exist then they won't know about other agents.

Doubt Question ⇒ These are still in the Markovian Domain. What about the Non-Markovian? Ans ⇒ Non Markovian / Non-Markovian Dynamic still can be compiled into MDPs.

There are still many other types ⇒ SMDP's , MMDP's

SMDP's ⇓ Time syn.

MMDP's ⇓ Multiple Agents having apriori information

* Finding optimal policy in MDP is polynomial in state level only.

Non-trivial learning still could be done at atomic level.

Case in Point :- Atomic is there is no structure. Even Linear Regression has structure to it.

$\boxed{\text{RL}}$ ⇒ Can be done at atomic level.

↳ Basically if I am in at a state then to avoid falling into bad decisions.

$\boxed{A^* \text{ Search}}$ ⇐ MDP differs

Scanned by CamScanner

Introduction to Gridworld

Ideas from Russell Norvig

* The transitions are Markovian in a sense that probability of reaching a state s' from s depends on only s & not on the history.
* Rewards must be "bounded".

MDP formally is

A sequential decision problem for a fully observable, stochastic environment with a Markovian transition model and additive rewards is called a MDP & consists of a set of states, set of Actions in each state, a transition model $P(s'|s,a)$ & reward f^n $R(s)$.

Policy

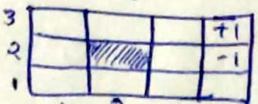
↳ A solution must specify what the agent should do for any state that the agent might reach. { Think back to DFA in ToC }

Optimal Policy → Policy that yields the highest expected utility.

Finite or Infinite Horizon

↳ After a fixed time N after which nothing matters

For ex start at (3,1) & N=3 then head directly.



If N=100 ⇒ take safer route from left.

With finite horizon → the optimal action in a given state could change over time.

1) Additive Rewards :- Utility = $R(s_0) + R(s_1) \ldots$

2) Discount Rewards :- $R(s_0) + \gamma R(s_1) + \text{---}$

Let's see non-terminal

Now our policy depends on the the $R(s)$.

Different $R(s)$ ⇒ Different behaviour

Moving forward we see that infinite horizon pose a natural question ⇒ 1) If agent does not reach a terminal state or if the agent never reaches one then ⇒ environment ∞

Utilities ∞

BUT

Infinite summation of Discount $= R_{max}/(1-\gamma)$

Points to note

* If agent agents up in a terminal state ⇒ no infinite sequences ↳ "Proper Policy".
* A way to deal with infinite seq would be the averaging argument.

Bottom Line

↳ Discounted Rewards present the fewest difficulties in evaluating state sequences.

Optimal Policy and Utilities of States

Recap

Utility ⇒ Sum of discounted rewards during the sequence

Compare Policies ??? ⇒ By Expected Utilities obtained when executing

Start

↳ Assume an agent is in some state s & define S_t {a variable} ⇓ a state agent reaches at time t. [$S_0 \equiv s$; simple enough]

Probability distribution over state sequence $S_1, S_2, \ldots$ is determined by the initial state s, the policy π & transition model for environment

Expected Utility $U^{\pi}(s) = E\left[\sum_{t=0}^{\infty} r^t R(S_t)\right]$

{Logical Reasoning :⇒ $E(x) = p * R(p)$ }

$\pi_s^* \to \max U^{\pi}(s)$

Basically although optimal, this optimality is based on the start state. {ie let's say i start from another state ⇒ other policy will be optimal; quite obvious.}

Now in case of infinite horizon optimal policy is independent of the starting state.

Short Proof

start with a & b

$a \to c \leftarrow b$

Intermediate c

↓ won't differ ∴ won't matter

POMDP's

Even the partially observ... is different from value parameterized f^n that we looked before. Even then the agent did ~~not~~ know about it's current location. We have some partial informa... about the states.