# Markov Decision Processes

Thursday, November 21, 2019 2:05 PM

- Incorporate Time to a decision network

observe → acts → reward / punishment

- long but unknown amount of time.
- indefinitely (infinite horizon process)

## Problem #1: Utility

Consider: A: \$1,000, \$1,000, \$1000, .....
B: \$1, \$1, \$1, \$1, ...
C: \$1,000,000, \$0, \$0, \$0, \$0, \$0, .....
D: \$0, \$0, \$0, ... \$1,000,000, \$0, \$0, \$0...

"Value"

- $V = \sum_{i=0}^{\infty} r_i$
- $V = \lim_{n \to \infty} (r_1 + \dots r_n)/n$
- $V = r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4 + \gamma^4 r_5 \dots$

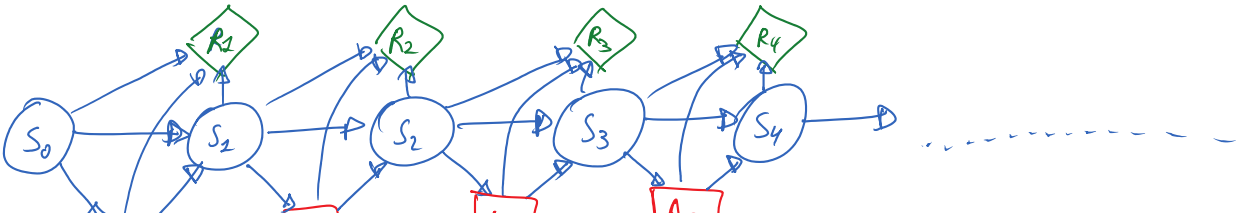
$\gamma$: discount rate $0 \le \gamma \le 1$

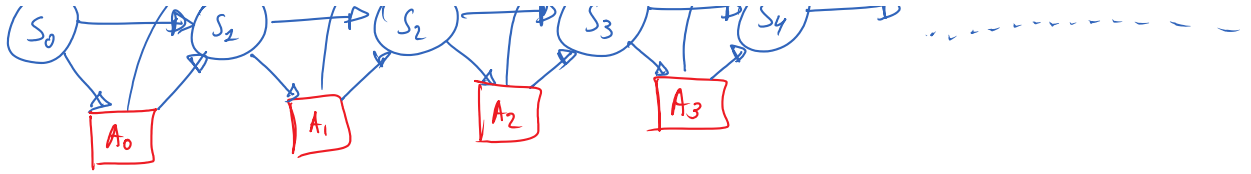
$= r_1 + \gamma(r_2 + \gamma(r_3 + \gamma(r_4 \dots)))$

$V_t = r_1 + \gamma V_{t+1}$

If we have the following properties
- The next state depends only on the current action.
- The effects of actions are stochastic, but do not change over time

## Belief Network: Markov Decision Process (MDP)






an MDP consists of:

- a set $S$ of states
- a set $A$ of actions
- the dynamics $P(S_{i+1} \mid S_i, A_i)$
- Reward table $R(s, a, s')$

  sometimes we use expected value of R $\quad R(s,a) = \sum_{s'} R(s,a,s') \cdot P(s' \mid s,a)$

- $\gamma$ the discount factor

Example:

Agent Bob $\quad S = \{\text{healthy}, \text{sick}\}$

$A = \{\text{relax}, \text{party}\}$

what should bob do each weekend?

$P(s' \mid s, a)$

| S | A | $P(s' = \text{healthy})$ |
|---|---|---|
| healthy | relax | 0.95 |
| healthy | party | 0.70 |
| sick | relax | 0.50 |
| sick | party | 0.10 |

$R(S, A)$

| S | A | |
|---|---|---|
| healthy | relax | 70 |
| healthy | party | 100 |
| sick | relax | 0 |
| sick | party | 20 |

- A Policy:

A policy $\pi : S \to A$ what to do at each state.

Among policies. $\pi^*$ Optimal policy is the one with maximum expected reward.

An MDP with stationary dynamics always has an optimal policy.

expected reward.

An MDP with stationary dynamics always has an optimal policy.

- Partially Observable Markov Decision Processes (POMDP)

  Markov Decision Process on a Hidden Markov Model