What is an MDP?

Optimization Problem $\quad$ maximize $E\left[\sum_{t=0}^{\infty} R(s_t, a_t)\right]$

Defined by (S, A, T, R, γ)

What is a policy?

- Closed loop, <u>deterministic</u> $\quad \pi : S \to A$
  "policy"

- Open loop, <u>deterministic</u>: $\quad$ List of actions executed in order

For deterministic problems
$\exists$ optimal open loop policy

Finite Horizon T

O.L. $\quad |A|^T$

C.L. $\quad |A|^{|S|^T}$

<u>Today</u>

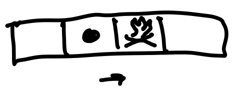
What is a Value Function?

How can we find optimal policies?

Dynamic Programming

Policy Iteration

$\pi(s) = \underset{a \in A}{\operatorname{argmax}}\ R(s,a)$ $\quad$ "Myopic" / "Greedy"

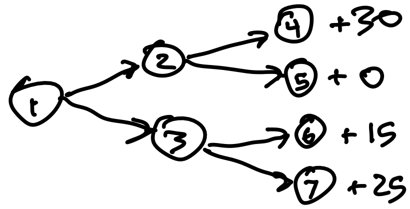


$+1$ for $\rightarrow$

-10 for

Dynamic Programming ~~Bre~~ Large problem → smaller sub

U/D



~~the~~ Naive

consider all possible plans

$R(U) + R(U) = 30 \leftarrow$

$U \quad D = 0$

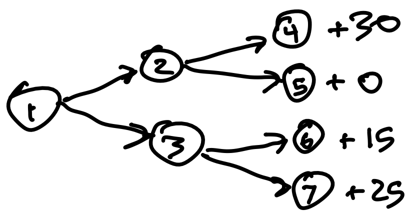
$D \quad U = 15$

$D \quad D = 25$

Value / Utility Function

$$V^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s, a_t = \pi(s_t)\right]$$

$$\pi^* = \arg\max V^{\pi}(s)$$



DP

$V(4) = 30$

$V(5) = 0$

$V(6) = 15$

$V(7) = 25$

3 computation steps
$\left\{ \begin{array}{l} \text{- at 2} \\ \text{at 3} \\ \text{at 1} \end{array} \right.$

$\pi^*(2) = U \qquad V^*(2) = 30$

$\pi^*(3) = D \qquad V^*(3) = 25$

$\pi^*(1) = U \qquad V^*(1) = 30$

Two Basic Algorithms

- Policy Iteration ← Easier to Understand
- Value Iteration ← Easier to Implement

Bellman's principle of optimality

Every sub-path of an optimal path is optimal

$$V^\pi(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s, a_t = \pi(s_t)\right]$$

$$= R(s, \pi(s)) + E\left[\sum_{t=1}^{\infty} \gamma^t R(s_t, a_t) \mid s_1 \sim T(s,a), a_t = \pi(s_t)\right]$$

$$= R(s, \pi(s)) + \gamma E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 \sim T(s,a), a_t = \pi(s_t)\right]$$

$$V^\pi(s) = R(s, \pi(s)) + \gamma E\left[V^\pi(s') \mid s' \sim T(s,a)\right]$$

$$\underset{s' \sim T(s,a)}{E}\left[V^\pi(s')\right]$$

Discrete state/action spaces

$$V^\pi(s) = R(s, \pi(s)) + \gamma \sum_{s' \in S} T(s' \mid s, \pi(s))\, V^\pi(s')$$

$$\underbrace{V^\pi}_{|S| \text{ vector}} = \underbrace{R^\pi}_{|S| \text{ vector}} + \gamma \underbrace{T^\pi}_{|S| \times |S| \text{ matrix}} \underbrace{V^\pi}_{|S| \text{ vector}}$$

$$T^\pi_{ij} = T^\pi(s' = j \mid s = i, \pi(i))$$

$|S|$ = number of element in S

$$V^\pi - \gamma T^\pi V^\pi = R^\pi$$

$$(I - \gamma T^\pi) V^\pi = R^\pi$$

$$V^\pi = (I - \gamma T^\pi)^{-1} R^\pi$$

← exact policy evaluation

$O(|S|^3)$ time

Policy iteration

π_0 = guess, $k = 0$

while $\pi_k \neq \pi_{k+1}$

$\quad V^{\pi_k} = (I - \gamma T^{\pi_k})^{-1} R^{\pi_k}$

$\quad$ for $s \in S$

$$\pi_{k+1}(s) = \underset{a \in A}{\operatorname{argmax}} \left(\overbrace{R(s,a) + \gamma \sum_{s' \in S} T(s' \mid s, a)\, V^{\pi_k}(s')}^{Q^\pi(s,a)}\right)$$

$\quad k = k + 1$

return π_k

optim

+10		+10
1	2	3

$\pi_a(2) = \leftarrow$ $\quad V^{\pi_a}(2) = \gamma 10$

$\pi_b(2) = \rightarrow$ $\quad V^{\pi_b}(2) = \gamma 10$

*

+8		+9

$$V^*(s) = \max_{a \in A} \left(R(s,a) + \gamma E\left[V^*(s') \mid s' \sim T(s,a)\right]\right)$$

Bellman's Equation

$$V^{k+1}(s) = \max_{a \in A} \left(R(s,a) + \gamma E\left[V^k(s') \mid s' \sim T(s,a)\right]\right)$$ repeat