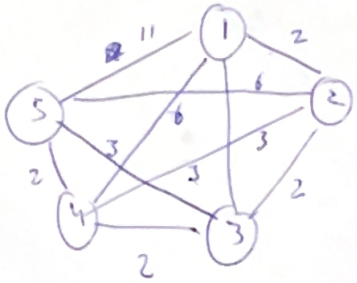
Rest DP RL Feb '23

(a) $$V_{n+1}(x) = \begin{cases} \min_y \{ d(x,y) + V_n(y) \} & x \neq \text{destination} \\ 0 & x = \text{destination} \end{cases}$$

$$V_0(x) = \begin{cases} \infty & x \neq \text{dest.} \\ 0 & x = \text{dest.} \end{cases}$$

(b)



(c)

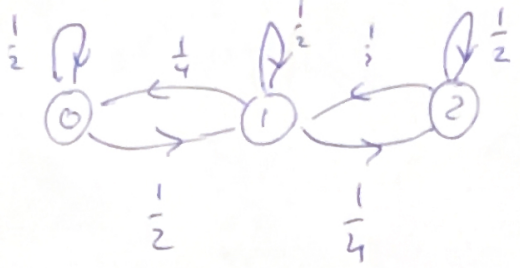
stage	1	2	3	4
1	6	11	∞	
2	5	5	6	∞
3	3	3	3	∞
4	2	2	2	∞
5	0	0	0	0
time	3	2	1	0

← calculation

shortest path: $1 \rightarrow 3 \rightarrow 5$

2a

$X = \{0, 1, 2\}$ = # machines up



b) $\pi(0) = \frac{1}{2}\pi(0) + \frac{1}{4}\pi(1)$

$\pi(1) = \frac{1}{2}\pi(1) + \frac{1}{2}\pi(0) + \frac{1}{2}\pi(2)$

$\pi(2) = \frac{1}{2}\pi(2) + \frac{1}{4}\pi(1)$

solution: $\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$.

c) $\phi^* = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)\begin{pmatrix}10\\1\\0\end{pmatrix} = \frac{10+2+1}{4} = \frac{13}{4}$

d) $V(0) + \phi = 10 + \frac{1}{2}V(0) + \frac{1}{2}V(1)$

$V(1) + \phi = 1 + \frac{1}{4}V(0) + \frac{1}{2}V(1) + \frac{1}{4}V(2)$

$V(2) + \phi = 0 + \frac{1}{2}V(1) + \frac{1}{2}V(2)$

We know $\phi = \frac{13}{4}$, set $V(0) = 0 \Rightarrow V(1) = -\frac{31}{2}$

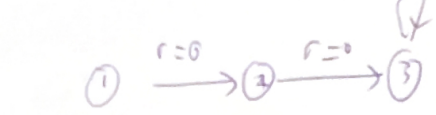
$\Rightarrow V(2) = -22$

e) $V(0) + \phi = \min\left\{10 + \frac{1}{2}V(0) + \frac{1}{2}V(1),\ 12 + \frac{1}{3}V(0) + \frac{2}{3}V(1)\right\}$

$V(1) + \phi = \min\left\{1 + \frac{1}{4}V(0) + \frac{1}{2}V(1) + \frac{1}{4}V(2),\ 2 + \frac{1}{6}V(0) + \frac{1}{2}V(1) + \frac{1}{3}V(2)\right\}$

$V(2) + \phi = \frac{1}{2}V(1) + \frac{1}{2}V(2)$

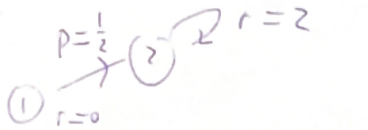
3a) $$m(i) = \max_T \frac{\mathbb{E}\sum_{t=1}^{T} \beta^t r(X_t \mid X_1 = i)}{\mathbb{E}\sum_{t=1}^{T} \beta^t}$$ with T a stopping time

b) 

$$m(1) = \frac{0 + \beta 0 + \beta^2 1 + \beta^3 1 + \cdots}{1 + \beta + \beta^2 + \cdots} = \beta^2$$

$$m(2) = \beta$$

$$m(3) = 1$$

c) 

$$V_1(1) = \frac{1}{2} V_1(2) + \frac{1}{2} V_1(3)$$

$$V_1(2) = \max\left\{2 + \frac{1}{2} V_1(2), V_1(1)\right\}$$

$$V_1(3) = \max\left\{\frac{1}{2} V_1(3), V_1(1)\right\}$$

$V_1(2) = 4$, $V_1(1) = \frac{4}{3} = V_1(3) \Rightarrow m(1) = (1-\beta)\frac{4}{3} = \frac{2}{3}$

d) Pull the second arm once (highest GI), when state 2 is reached continue forever, otherwise switch

e) exp discounted rewards $= 0 + \frac{1}{2}\left(\frac{1}{2}\left(2 + \frac{1}{2}2 + \left(\frac{1}{2}\right)^2 + \cdots\right) + \frac{1}{2}\left(0 + \frac{1}{2}0 + \left(\frac{1}{2}\right)^2 1 + \left(\frac{1}{2}\right)^3 1 + \cdots\right)\right)$

$$= \frac{1}{2}\left(2 + \left(\frac{1}{2}\right)^3 2\right) = 1 + \frac{1}{8} = \frac{9}{8}$$

4a. Cite 3 elements that makes the RL setting general in the sense that it can represent basically all sequential decision-making problems (tip: think about the simplest MDP possible and all the elements that can make it more complex)
A: 3 out of these: large state/action space, partial observability, stochastic transition, finite data available or access to an inaccurate simulator.
b. Why is it not possible to solve the game of chess with simple MCTS algorithms (without any value function estimator)?
A: the depth and breadth of search is too big
c. The Bellman equation that is at the core of reinforcement learning makes use of the fact that the Q-function can be written in a recursive form. Can you write this Bellman equation? (Make sure that you define all the terms).
A: $Q(x,a) = r(x,a) + \beta \sum_y p(y|x,a) \max'_a Q(y,a')$
d. Describe in words the "optimistic values" exploration strategy.
A: by setting the initial Q-values high actions are selected until their value is decreased to what is likely to be their real value.
e. What is the reinforce algorithm?
A: It estimates $Q^\pi(x,a)$ from (on-policy) rollouts on the environment while following policy π.