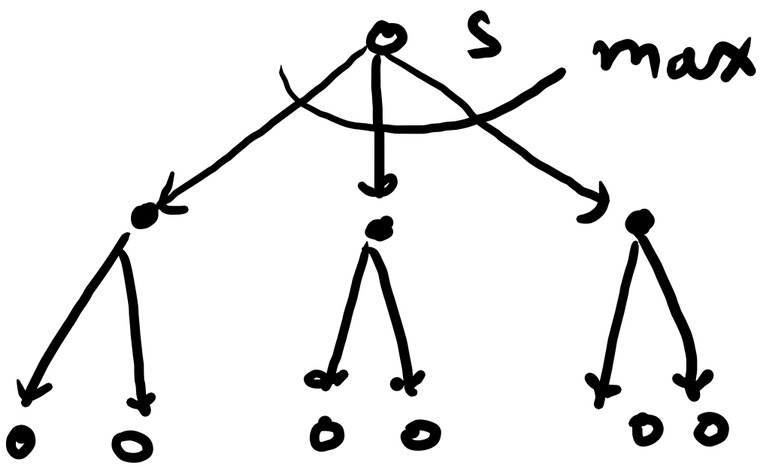
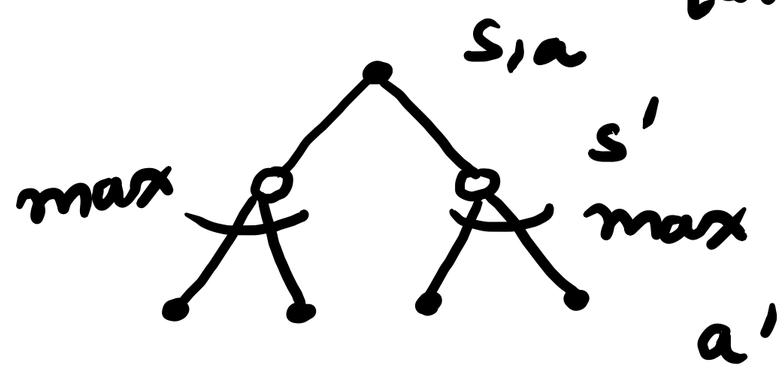
Lec-10, IT567, 24-25

We have derived or seen the Bellman optimality equation (BOE) in the last lecture. *

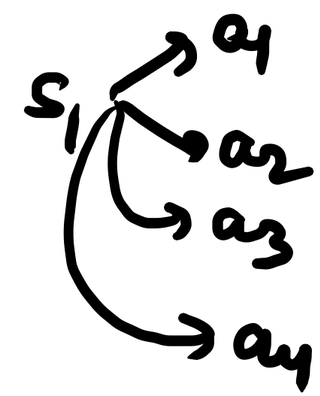
Let's draw the back up diagrams for BOE

* $$V_*(s) = \max_a \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V_*(s') \right]$$





$q_*(s, a)$ $\quad s \to s_1$



$$q_*(s,a) = E\left[R_{t+1} + \gamma V_*(S_{t+1}) \mid S_t = s, A_t = a\right]$$

$$= E\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a\right]$$

$$\hookrightarrow \sum_{s'} \sum_{r} p(s', r \mid s, a)\left[r + \gamma \cdot \max_{a'} q_*(s', a')\right]$$

$$\because V_*(s) = \max_{a} q_*(s,a)$$

The backup diagrams are same except that arcs have been added at the agent's choice points to represent that the max over that choice is taken rather than the expected value given some value.

For finite MDPs, BOE has a unique solution. BOE is actually a system of equations, one for each state — n states there are n equations in n unknowns. If dynamics ($p(s', r|s, a)$) of the environment is known, solve system of equations for V_π using a method for solving system of "non-linear" equations

$$\begin{bmatrix} 2 & 3 & 4 \\ 6 & 7 & 8 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 2 \\ 0 \end{bmatrix}$$

$$2x_1 + 3x_2 + 4x_3 = 7$$
$$6x_1 + 7x_2 + 8x_3 = 2$$
$$-x_1 + x_3 = 0$$

system of linear equations

Chapter 4: Dynamic programming (DP)

DP refers to a collection of algorithms that can be used to compute optimal policies given a perfect model of the environment as MDP. (Model-Based)

issues → model availability
→ great computational expense

We will study model-based methods initially but methods after are attempts to achieve the same effect as DP only with less computation & without assuming a perfect model of the environment.

DP algorithms are obtained by turning BE into assign-ments, i.e., update rules for improving approx. of the desired value functions.