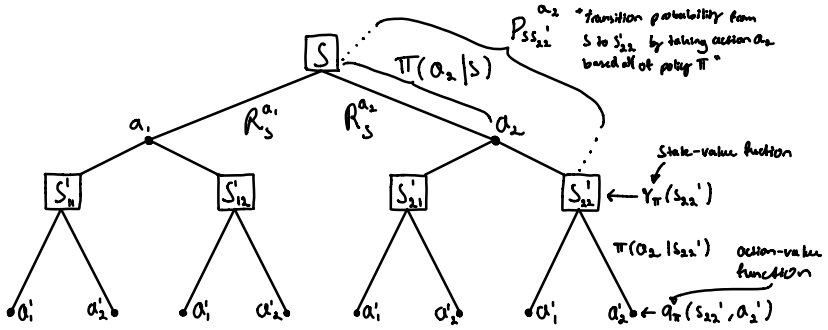
Markov Decision Process: $\langle S, A, P^{a}_{ss'}, R^{a}_{s}, \gamma \rangle$

graph



$$v_\pi(s) = \sum_{a \in A} \pi(a|s) \left[\mathcal{R}^{a}_{s} + \gamma \sum_{s' \in S} P^{a}_{ss'} v_\pi(s') \right]$$

$$q_\pi(s,a) = \mathcal{R}^{a}_{s} + \gamma \sum_{s' \in S} P^{a}_{ss'} \cdot \sum_{a' \in A} \pi(a'|s') \cdot q_\pi(s', a')$$

ex:

$$v_\pi(s) = \pi(a_1|s) \cdot \left[\mathcal{R}_s^{a_1} + \gamma \left[P^{a_1}_{ss'_{11}} \cdot v_\pi(s'_{11}) + P^{a_1}_{ss'_{12}} \cdot v_\pi(s'_{12}) \right] \right] +$$
$$\pi(a_2|s) \cdot \left[\mathcal{R}_s^{a_2} + \gamma \left[P^{a_2}_{ss'_{21}} \cdot v_\pi(s'_{21}) + P^{a_2}_{ss'_{22}} \cdot v_\pi(s'_{22}) \right] \right]$$

probability we go from $s \to s'_{22}$ w/ a_2 multiplied by the value of s'_{22} — the value of the state scaled by the likelihood we end up there

discounted sum of value of branches scaled by the likelihood we end up at one of the branches

value of immediate transition by a_2, plus averaged value of the states we would end up in given that we took a_2 at s

sum of probability we end up at a state multiplied by its value

* the whole thing:

The value of a state is defined as the average value of taking

starting bounds reel: [going from $V \to V_\pi$]

"in terms of" $q \to v$

1b.

$$q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \cdot v_\pi(s')$$

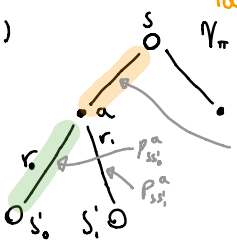
the value of taking an action, a, at a state, s, under policy, π, (represented by $q_\pi(\cdot,\cdot)$) is the reward for taking the action, R_s^a, which is given by the environment, in addition to the discounted [γ] average value of the states you would end up at.

determined by env.

odds we'll end up there · value

· high value · high odds ⇒ big influence on q_π

· low value · high odds ⇒



"in terms of" $v \to q$

1a.

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) \cdot q_\pi(s,a)$$

$\pi(a|s)$

↳ assume a stationary policy, one that doesn't change throughout an episode

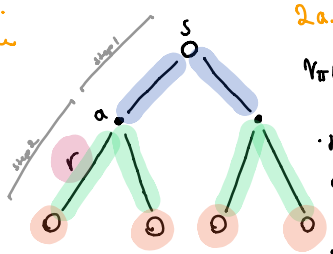
Value, $v_\pi(\cdot)$ of state, s, under policy, π, is the average value of the actions, $q_\pi(s,\cdot)$ available

the intuition is that an action leads you to another state, which itself represents the future income sense

↳ a state is valuable if it leads you to more valuable states

- By averaging, we take into account probability we will pick an action/odds we'll end up somewhere as well as the value of the result ↘
- Both of these definitions are mindful of the fact that taking an action does not translate to going to another state.
- this notion also explains why reward, R_s^a, comes from the result of taking a, not the intent of taking a

1a. & 1b. define quantities by looking a single step ahead (1 level in the graph). By looking further we can form recursive definitions.



2a.

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) \underbrace{\left[R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \cdot v_\pi(s') \right]}_{q_\pi(s,\cdot) =}$$

- the prior intuition is enforced: by looking at the average values of the possible actions available we judge how valuable a state is, and the value of one of the averaged actions is the sum of the immediate reward, and the average value of the resultant states (includes entropy from env.)

"in terms of" $q \to q'$

2b.

$$q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \underbrace{\left[\sum \pi(a'|s') q_\pi(s',a') \right]}_{v_\pi(s')}$$

"in terms of" $v \to v'$

- the value of an action available at state s is the sum of the immediate reward, and the discounted average of the states we'll end up at, which is represented as the average value of the actions available at those future states

we value an action by its immediate reward and the various actions we'll take as a consequence of the current one (which includes the rewards of future actions)

1a./b. lead us to 2a./b. which describe methods for navigating an MDP under policy π. Its practical to figure out an optimal policy, π^* for solving problems posed as MDPs.

least possible performance in an MDP

$$v_*(s) = \max_\pi v_\pi(s)$$ ← the optimal state-value function, $v_*(\cdot)$, is the one which is the highest value across all possible policies

$$q_*(s,a) = \max_\pi q_\pi(s,a)$$ ← the optimal action-value function, q_*, is the best action-value across all policies

scalar comparison

with these definitions, can optimize π w/ q_* & v_* as objective functions

↳ assume $\pi \geq \pi'$ if $v_\pi(s) \geq v_{\pi'}(s)$

- there exists optimal policy, $\pi_* \geq \pi$, $\forall \pi$
- for clarity, $v_{\pi_*}(s) = v_*(s)$ & $q_{\pi_*}(s,a) = q_*(s,a)$

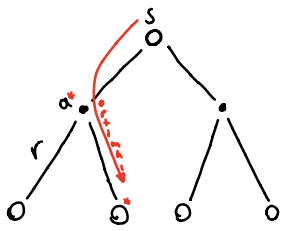
since you've always made the right choice

3b.

$$q_*(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \, v_*(s')$$

$$= R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \left[\max_{a'} q_*(s', a') \right]$$

constrain to make best local selections



3a.

$$v_*(s) = \max_a q_*(s,a)$$

$$= \max_a \left[R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \, v_*(s') \right]$$

The addition of "optimal" (3a./3b.) only modifies the before/look ahead equations, and stops them being their optimal form.

↳ this aligns w/ the idea that for an optimal outcome we need to make a locally good choice, and assume we'll make a good choice in the future too.

Bellman equality is non-linear

↳ these are the basis for
- value iteration
- policy iteration
- Q-learning
- Sarsa