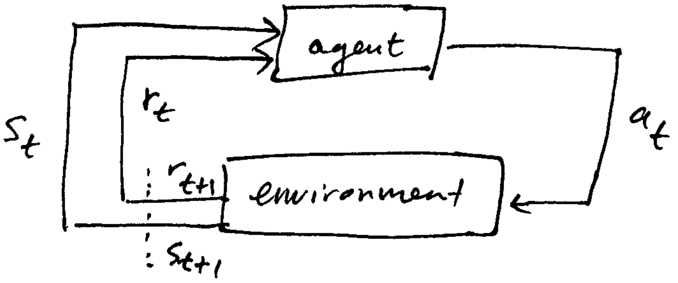
MARKOV DECISION PROCESSES



A <u>Markov</u> <u>decision</u> <u>process</u> satisfies

$$\Pr\{s_{t+1}=s',\ r_{t+1}=r \mid s_t, a_t, \ldots, s_0, a_0\}$$

$$= \Pr\{s_{t+1}=s',\ r_{t+1}=r \mid s_t, a_t\}$$

That is, the state and reward at time $t+1$ depend only on the state and action at the previous step, and not on the more distant past.

Many RL problem can be cast (at least approximately) as MDPs. Note that we get to define the state space, and we can choose a state space such that the Markov property holds.

Example Autonomous navigation in 2d: $s_t = [x(t)\ \ y(t)\ \ \dot{x}(t)\ \ \dot{y}(t)\ \ \ddot{x}(t)\ \ \ddot{y}(t)]^T$

By def. of conditional probability,

$$\Pr\{s_{t+1} = s', r_{t+1} = r \mid s_t = s, a_t = a\} =$$

$$\Pr\{s_{t+1} = s' \mid s_t = s, a_t = a\}$$

$$\times \Pr\{r_{t+1} = r \mid s_{t+1} = s', s_t = s, a_t = a\}.$$

It is therefore common to characterize an MDP by the transition probabilities

Ⓐ $$P^a_{ss'} :=$$

and by

$$R^a_{ss'} :=$$

The Bellman Equations

Let $\pi = \pi(s,a)$ be a policy. Recall

$$V^\pi(s) = E_\pi\{R_t \mid s_t = s\}$$

$$Q^\pi(s,a) = E_\pi\{R_t \mid s_t = s, a_t = a\}$$

are the state and state-action value functions.

Unless otherwise stated, we will assume a finite MDP, i.e. $|S| < \infty$, $|A| < \infty$, $S=$ state space, $A=$ action space

$$\begin{aligned}
V^{\pi}(s) &= E_{\pi}\{R_t \mid s_t = s\} \\
&= \sum_{a} \pi(s,a)\, E_{\pi}\{R_t \mid s_t = s,\ a_t = a\} \\
&= \sum_{a} \pi(s,a) \cdot \sum_{s'} P^{a}_{ss'}\, E_{\pi}\{R_t \mid s_{t+1} = s',\ s_t = s,\ a_t = a\}
\end{aligned}$$

$$\left[\text{write } R_t = \sum_{k\geq 0} \gamma^k r_{t+k+1} = r_{t+1} + \gamma \sum_{k\geq 0} \gamma^k r_{t+k+2} = r_{t+1} + \gamma \cdot R_{t+1}\right]$$

$$\begin{aligned}
&= \sum_{a,s'} \pi(s,a) P^{a}_{ss'}\, E_{\pi}\{r_{t+1} + \gamma \cdot R_{t+1} \mid s_{t+1} = s',\ s_t = s,\ a_t = a\} \\
&= \sum_{a,s'} \pi(s,a) P^{a}_{ss'} \left[R^{a}_{ss'} + \gamma V^{\pi}(s')\right]
\end{aligned}$$

This gives a system of equations (linear!) in the values $V^{\pi}(s)$. Therefore, given any policy, and knowledge of the MDP dynamics ($P^{a}_{ss'}$, $R^{a}_{ss'}$), we can determine the value function of π by solving a linear system of equations.

Similarly, for the state-action value function, we have

(B) $Q^{\pi}(s,a) =$

Unfortunately, for many problems, the state space S is too large for it to be practical to solve the Bellman equations directly. E.g., for backgammon, $|S| \approx 10^{20}$.

Nonetheless, these equations will form the basis of more efficient algorithms.

It can be shown that the solution to the Bellman equations for V^{π} exists and is unique.

Optimal Policies

We say $\pi \geq \pi' \iff V^{\pi}(s) \geq V^{\pi'}(s) \quad \forall s$

We say π^* is optimal iff $\pi^* \geq \pi$ for all π.

There is always an optimal policy. There may be more than one.

Even if there are many optimal policies, they all have the same value-function the optimal state-value function:

$$V^*(s) = \max_{\pi} V^{\pi}(s), \qquad s \in S$$

and the optimal action-value function

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) \qquad s \in S, a \in A$$

= expected return for taking action a in state s, and thereafter following an optimal policy

=

©

Note that the final equation does not involve π^*!

Using this fact, we can write

$$V^*(s) = \max_a Q^*(s,a)$$

$$= \max_a \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V^*(s') \right]$$

These are the Bellman optimality equations. They are a nonlinear system of equations.

Many RL algorithms can be understood as (approximately) solving the Bellman optimality equations.

Once V^* or Q^* are known, an optimal policy can be determined. If Q^* is known:

Ⓓ $$\pi^*(s) =$$

If V^* is known

$$\pi^*(s) =$$

Key

A. $P^a_{ss'} = Pr\{s_{t+1} = s' \mid s_t = s, a_t = a\}$

$R^a_{ss'} = E\{r_{t+1} \mid s_t = s, a_t = a, s_{t+1} = s'\}$

B. $Q^\pi(s,a) = E\{R_t \mid s_t = s, a_t = a\}$

$$= \sum_{s'} P^a_{ss'} E_\pi \left[r_{t+1} + \gamma R_{t+1} / s_t = s, a_t = a, s_{t+1} = s' \right]$$

$$= \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V^\pi(s') \right]$$

$$= \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma \sum_{a'} \pi(s',a') Q^\pi(s',a') \right]$$

C. $Q^*(s,a) = Q^{\pi^*}(s,a)$

$$= \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V^{\pi^*}(s') \right]$$

$$= \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V^*(s') \right]$$

$$= \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma \max_{a'} Q^*(s,a') \right]$$

D. $\pi^*(s) =$ any solution of $\max_a Q^*(s,a)$

$\pi^*(s) =$ any solution of $\max_a \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V^*(s') \right]$