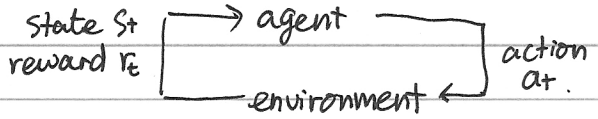
Review

* Reinforcement Learning.



* Markov Decision Process. (MDP).

$\{S, A, P(s'|s,a), R(s)\}$.

states, actions, transition probabilities, reward.

* Policy : assignment of states to actions $\pi(s) \in A$.

* State Value Function.

$$V^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \mid s_t = s\right]$$ "expected value of discounted reward"

$$V^{\pi}(s) = R(s) + \gamma \sum_{s'} P(s'|s, \pi(s)) \cdot V^{\pi}(s')$$ Bellman Equation.

* Action value Function.

$Q^{\pi}(s,a)$ = "expected return from initial state s, taking action a, then following policy π"

$$= \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \cdot R(s_t) \mid s_0 = s, a_0 = a\right]$$

$$= R(s) + \gamma \cdot \sum_{s'} P(s'|s,a) \cdot V^{\pi}(s')$$

* Optimality

Theorem: there is always at least one policy π^* for which $V^{\pi^*}(s) \geq V^{\pi}(s)$ for all s, π.

* Optimal value functions

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

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

- There may be many (equivalently) optimal policies, but optimal value functions (both state and action) are unique.

* Relations between value functions and policies

- Given MDP $\{S, A, P(s'|s,a), R(s), \gamma\}$ and given $\pi^*(s)$. then it is easy to write out $V^*(s)$ and $Q^*(s,a)$.

$$V^*(s) = R(s) + \gamma \cdot \sum_{s'} P(s'|s, \pi^*(s)) \cdot V^*(s')$$

$$Q^*(s,a) = R(s) + \gamma \cdot \sum_{s'} P(s'|s,a) \cdot V^*(s')$$

- Vice versa. (from optimal value functions to optimal policies).

$$\pi^*(s) = \underset{a}{\text{argmax}}\ Q^*(s,a)$$

$$= \underset{a}{\text{argmax}} \left[R(s) + \gamma \cdot \sum_{s'} P(s'|s,a) \cdot V^*(s') \right]$$

$$= \underset{a}{\text{argmax}} \left[\sum_{s'} P(s'|s,a) \cdot V^*(s') \right]$$

* **Planning under Uncertainty.**

Assume complete model of environment as.

$MDP = \{S, A, P(s'|s,a), R(s), 0 \le \gamma \le 1\}$.

how to compute $\pi^*(s)$, or equivalently, $V^*(s)$ or $Q^*(s,a)$?

1). Policy Evaluation. - how to compute $V^\pi(s)$ for any (possibly non-optimal) policy π?

From Bellman equation:

$$V^\pi(s) = R(s) + \gamma \cdot \sum_{s'=1}^{n} P(s'|s, \pi(s)) \cdot V^\pi(s')$$

(given) for $s = 1, 2, 3, \ldots n$. where n = # states in S.

This is a system of n linear equation for n unknowns.

Put all unknowns on LHS.

$$V^\pi(s) - \gamma \cdot \sum_{s'} P(s'|s, \pi(s)) \cdot V^\pi(s') = R(s)$$

$$\sum_{s'} \left[I(s,s') - \gamma \cdot P(s'|s, \pi(s)) \right] \cdot V^\pi(s') = R(s) \quad \text{for } s = 1, 2, \ldots, n.$$

- Rewrite equation as:

$$(I - \gamma \cdot P)\, V = R$$

I: $n \times n$ identity matrix; V: unknown $n \times 1$ vector; R: known $n \times 1$ vector.

matrix is always invertible for $0 \le \gamma < 1$.

Solution: $V^{\pi} = (I - \gamma P_{\pi})^{-1} \cdot R$

- Matrix inversion is $O(n^3)$ operation.

Ex: States $S \in \{0, 1\}$.

Transitions $p^{\pi}(s' | s, \pi(s)) = \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix}$ "rows sum to one"

Rewards $R(s) = \begin{bmatrix} r_0 \\ r_1 \end{bmatrix}$

State value function $V^{\pi}(s) = \begin{pmatrix} V_0 \\ V_1 \end{pmatrix}$.

Solve: $\left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \gamma \cdot \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix} \right] \cdot \begin{pmatrix} V_0 \\ V_1 \end{pmatrix} = \begin{pmatrix} r_0 \\ r_1 \end{pmatrix}$ 2 equations for 2 unknowns.

2) Policy improvement.

* How to compute π' such that $V^{\pi'}(s) \ge V^{\pi}(s)$ for all states s?
* Recall $Q^{\pi}(s, a)$ "expected return from state s, follow action a, then follow policy π".

How to compute $Q^{\pi}(s, a)$?

- Evaluate policy to get $V^{\pi}(s)$.
- $Q^{\pi}(s, a) = R(s) + \gamma \cdot \sum_{s'} P(s' | s, a) \cdot V^{\pi}(s')$

Define "greedy policy":

$$\pi'(s) = \underset{a}{\operatorname{argmax}} \; Q^{\pi}(s, a)$$

$$= \underset{a}{\operatorname{argmax}} \left[\sum_{s'} P(s' | s, a) \cdot V^{\pi}(s') \right]$$

Theorem: greedy policy π' everywhere performs better or equal to original policy π.

$V^{\pi'}(s) \ge V^{\pi}(s)$ for all s.

Intuition: if better to choose action a in state s, then follow π, it's always better to choose action a in state s.

Proof: $V^{\pi}(s) = Q^{\pi}(s, \pi(s))$

$\leq \max_a Q^{\pi}(s, a)$

$= Q^{\pi}(s, \pi'(s))$ by definition of greedy policy.

$= R(s) + \gamma \cdot \sum_{s'} P(s' | s, \pi'(s)) \cdot V^{\pi}(s')$

So far, it is better to take one step under π', then revert to π, than to follow π.

"one-step inequality": $V^{\pi}(s) \leq R(s) + \gamma \cdot \sum_{s'} P(s' | s, \pi'(s)) \cdot V^{\pi}(s')$

Apply inequality to $V^{\pi}(s')$ on RHS. ↓

$V^{\pi}(s) \leq R(s) + \gamma \cdot \sum_{s'} P(s' | s, \pi'(s)) \cdot \left[R(s') + \gamma \cdot \sum_{s''} P(s'' | s', \pi'(s') \cdot V^{\pi}(s'') \right]$

So, Better to take two steps under π', then revert to π, than to always follow π.

In general, apply "one-step inequality" t times, we will show:
Better to take $t+1$ steps under π', then follow π, than to always follow π.

Let $t \to \infty$, it's always better to follow $\pi'(s)$ than $\pi(s)$.

$\Rightarrow V^{\pi}(s) \leq V^{\pi'}(s)$ since RHS converges to $V^{\pi'}(s)$ for $\gamma < 1$.