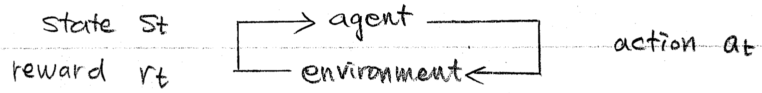
11/16 **Review**

* Reinforcement learning



* Markov decision process

$$MDP = \{ \underset{\text{states}}{\mathcal{S}}, \underset{\text{actions}}{\mathcal{A}}, \underset{\text{transitions}}{P(s'|s,a)}, \underset{\text{rewards}}{R(s)} \}$$

* How to learn from experience?

$$\begin{matrix} \text{state} & s_0 & \xrightarrow{\pi(s_0)} & s_1 & \xrightarrow{\pi(s_1)} & \cdots \\ \text{reward} & r_0 & & r_1 & & \end{matrix}$$

policy π

* State value function

$$V^{\pi}(s) = E^{\pi}\Big[\sum_{t=0}^{\infty} \gamma^t R(s_t) \,\Big|\, s_0 = s \Big]$$ expected long-term discounted return.

γ: discount factor $0 \le \gamma < 1$

Recursion relation:

$$\begin{aligned} V^{\pi}(s) &= E^{\pi}[R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots \mid s_0 = s] \\ &= R(s) + \gamma E^{\pi}[R(s_1) + \gamma R(s_2) + \gamma^2 R(s_3) + \cdots \mid s_0 = s] \\ &= R(s) + \gamma \sum_{s'} P(s'|s, \pi(s)) E^{\pi}[R(s_1) + \gamma R(s_2) + \cdots \mid s_1 = s'] \end{aligned}$$

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

* Action-value function

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

$$\begin{aligned} &= E^{\pi}\Big[\sum_{t=0}^{\infty} \gamma^t R(s_t) \,\Big|\, s_0 = s, a_0 = a \Big] \\ &= R(s) + \gamma \sum_{s'} P(s'|s,a) V^{\pi}(s') \end{aligned}$$

* Optimality

- optimal policy π^*

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

  Proof: by construction.

- optimal state-value function

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

- optimal action-value function

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

- relations

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

$$\pi^*(s) = \arg\max_a Q^*(s,a)$$
$$= \arg\max_a \sum_{s'} P(s'|s,a)\, V^*(s')$$

Planning

Assume complete model of environment MDP= $\{\mathcal{S}, \mathcal{A}, P(s'|s,a), R(s)\}$ and γ.

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

Q1. Policy evalution

How to compute $V^\pi(s)$ for fixed policy π?

From Bellman equation:

$$V^\pi(s) = R(s) + \gamma \sum_{s'} P(s'|s,\pi(s))\, V^\pi(s') \quad \text{for } s=1,2,\cdots,N$$

This is a system of N linear equations

$$V^\pi(s) - \gamma \sum_{s'} P(s'|s,\pi(s))\, V^\pi(s') = R(s)$$
$$\sum_{s'} \Big[I(s,s') - \gamma \underbrace{P(s'|s,\pi(s))}_{P^\pi}\Big] V^\pi(s') = R(s)$$
$$\underbrace{\left[I - \gamma P^\pi\right]}_{\text{known } n\times n \text{ matrix in terms of MDP}} \underbrace{V^\pi}_{\text{vector of unknowns}} = \underbrace{R}_{\text{vector of rewards}}$$

$$V^\pi = (I - \gamma P^\pi)^{-1} R \qquad \text{always invertible.}$$

Matrix inversion $O(N^3)$

HW: iterative solution of Bellman equations.

Q2. Policy improvement

How to compute π' such that $V^{\pi'}(s) \geq V^\pi(s)$ for all s?

* Define greedy policy $\pi'(s)$ by:

$$\pi'(s) = \arg\max_a \left[Q^\pi(s,a)\right]$$
$$= \arg\max_a \left[R(s) + \gamma \sum_{s'} P(s'|s,a)\, V^\pi(s')\right]$$
$$= \arg\max_a \left[\sum_{s'} P(s'|s,a)\, V^\pi(s')\right]$$

* Thm: greedy policy π' everywhere performs equally or better than original policy π: $V^{\pi'}(s) \geq V^\pi(s)$ for all s.

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

* Proof:
$$V^\pi(s) = Q^\pi(s,\pi(s))$$
$$\leq \max_a Q^\pi(s,a)$$
$$= Q^\pi(s,\pi'(s))$$

Expand: $V^\pi(s) \leq R(s) + \gamma \sum_{s'} P(s'|s,\pi'(s))\, V^\pi(s')$

So far, better to take one step under π' and revert to π, than to follow π.

Apply "one-step" inequality to $V^{\pi}(s')$ on RHS:

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

Better to take 2 steps under π' and revert to π than to always follow π.

Apply "t" times : better to take t steps under π' and revert to π
than to always follow π.

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

$\Longrightarrow V^{\pi}(s) \le V^{\pi'}(s)$. since RHS converges to $V^{\pi'}(s)$
assuming $\gamma < 1$.

Q3. Policy iteration

How to compute π^*?

Algorithm : 1) initialize policy at random
2) repeat until convergence
- compute value function
- derive greedy policy.

$$\pi_0 \xrightarrow{\text{evaluate}} \begin{matrix} V^{\pi_0}(s) \\ Q^{\pi_0}(s,a) \end{matrix} \xrightarrow{\text{improve}} \pi_1 \xrightarrow{\text{evaluate}} \cdots$$

* Is this guaranteed to converge? Yes.
- There are finite # policies $|A|^{|S|}$
- Policies cannot be indefinitely improved.
- Typically converge in far less steps than $|A|^{|S|}$.

* Does it always converge to optimal π^*? Yes!

Thm : suppose $\pi'(s) = \arg\max_a Q^{\pi}(s,a)$ and $V^{\pi'}(s) = V^{\pi}(s)$ for all s.
Then, $V^{\pi}(s) = V^{*}(s)$.
Note optimal value function $V^*(s)$ is unique even if there are many optimal policies.

* Proof strategy
1) Derive "Bellman optimality equation" satisfied by $V^{\pi}(s)$ when $V^{\pi}(s) = V^{\pi'}(s)$.
2) Show that $V^{\pi}(s) \ge V^{\hat{\pi}}(s)$ for all states and policies $\hat{\pi}$.
Hence $V^{\pi}(s) = V^{*}(s)$.

* Proof

1) From Bellman equation for $\pi'(s)$:

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

By assumption : $V^{\pi}(s) = V^{\pi'}(s)$ b/c we're converged

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

By assumption, π' is greedy with respect to $V^{\pi}(s)$:

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

⇒ Bellman optimality equation. set of <u>nonlinear</u> equation for $s = 1, 2, \cdots, n$

Different than linear Bellman equation for arbitrary policies.

* Why nonlinear?

$$\max_{a} f(a) = \lim_{\lambda \to \infty} \frac{1}{\lambda} \log \sum_{a} e^{\lambda f(a)}.$$