

# Statistical Reinforcement Learning

## - Note 1

### MDP

#### Value and Policy

Boundedness of **rewards** :  $r_t \in [0, R_{\max}]$

Boundedness of  $\mathbb{E} [\sum_{t=1}^{\infty} \gamma^{t-1} r_t]$  :  $\mathbb{E} [\sum_{t=1}^{\infty} \gamma^{t-1} r_t] \in [0, \frac{R_{\max}}{1-\gamma}]$

-Reason: 等比级数:  $\sum_{n=0}^{\infty} aq^n (a \neq 0)$

当  $0 < |q| < 1$  时,  $\sum_{n=0}^{\infty} aq^n$  收敛, 且

收敛于  $\frac{a}{1-q}$

Define  $V^{\pi}(s) = \mathbb{E} [\sum_{t=1}^{\infty} \gamma^{t-1} r_t | s_1 = s, \pi]$

So,  $V^{\pi}(s) < \frac{R_{\max}}{1-\gamma}$

#### Policy evaluation

Bellman equation for policy evaluation

$$\begin{aligned} V^{\pi}(s) &= \mathbb{E} \left[ \sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_1 = s, \pi \right] \\ &= \mathbb{E} \left[ r_1 + \sum_{t=2}^{\infty} \gamma^{t-1} r_t \mid s_1 = s, \pi \right] \\ &= R(s, \pi(s)) + \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s)) \mathbb{E} \left[ \gamma \sum_{t=2}^{\infty} \gamma^{t-2} r_t \mid s_1 = s, s_2 = s', \pi \right] \\ &= R(s, \pi(s)) + \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s)) \mathbb{E} \left[ \gamma \sum_{t=2}^{\infty} \gamma^{t-2} r_t \mid s_2 = s', \pi \right] \\ &= R(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s)) \mathbb{E} \left[ \sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_1 = s', \pi \right] \\ &= R(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s)) V^{\pi}(s') \\ &= R(s, \pi(s)) + \gamma \langle P(\cdot | s, \pi(s)), V^{\pi}(\cdot) \rangle \end{aligned}$$

## Bellman equation for policy evaluation

$$V^\pi(s) = R(s, \pi(s)) + \gamma \langle P(\cdot | s, \pi(s)), V^\pi(\cdot) \rangle$$

Matrix form: define

- $V^\pi$  as the  $|S| \times 1$  vector  $[V^\pi(s)]_{s \in S}$
- $R^\pi$  as the vector  $[R(s, \pi(s))]_{s \in S}$
- $P^\pi$  as the matrix  $[P(s' | s, \pi(s))]_{s \in S, s' \in S}$

$$V^\pi = R^\pi + \gamma P^\pi V^\pi$$

$$(I - \gamma P^\pi) V^\pi = R^\pi$$

$$V^\pi = (I - \gamma P^\pi)^{-1} R^\pi$$

This is always invertible. Proof?

20

## State occupancy

$$(I - \gamma P^\pi)^{-1}$$

Each row (indexed by  $s$ ) is the discounted state occupancy  $d_s^\pi$ , whose  $(s')$ -th entry is

$$d_s^\pi(s') = \mathbb{E} \left[ \sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{I}[s_t = s'] \mid s_1 = s, \pi \right]$$

- Each row is like a distribution vector—except that the entries sum up to  $1/(1-\gamma)$ . Let  $\eta_s^\pi = (1 - \gamma) d_s^\pi$  denote the normalized vector.
- $V^\pi(s)$  is the dot product between  $d_s^\pi$  and reward vector
- Can also be interpreted as the value function of indicator reward function

20

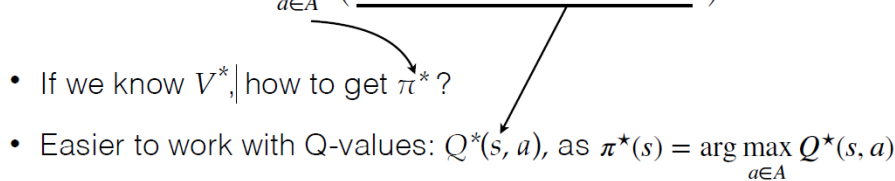
## Optimality

- For infinite-horizon discounted MDPs, there always exists a stationary and deterministic policy that is optimal for all starting states simultaneously

- Proof: Puterman'94, Thm 6.2.7 (reference due to Shipra Agrawal)

- Let  $\pi^*$  denote this optimal policy, and  $V^* := V^{\pi^*}$

- Bellman Optimality Equation:

$$V^*(s) = \max_{a \in A} \left( R(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} [V^*(s')] \right)$$


- If we know  $V^*$ , how to get  $\pi^*$ ?
- Easier to work with Q-values:  $Q^*(s, a)$ , as  $\pi^*(s) = \arg \max_{a \in A} Q^*(s, a)$

$$Q^*(s, a) = R(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} \left[ \max_{a' \in A} Q^*(s', a') \right]$$