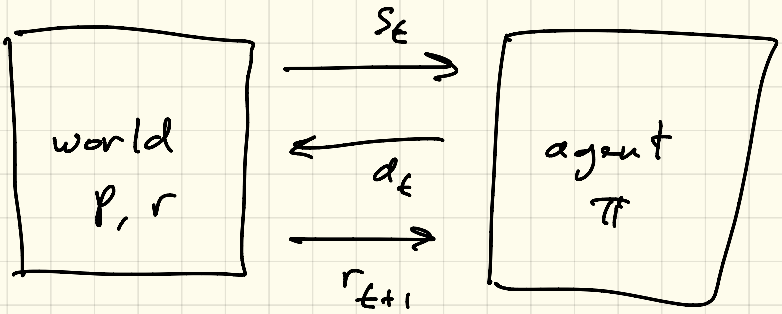
Value - function methods

MDPs — Markov Decision Process



t = time

spaces : s_t state, $\in \{1, \dots, N\}$, a_t action, $\in \{1, \dots, M\}$, r_{t+1} reward $\in \mathbb{R}$

$p: S \times S \times A \to [0,1]$

functions : model $p(s_{t+1} \mid s_t, a_t)$ = prob. next state is s_{t+1}, given s_t, a_t

$\pi: A \times S \to [0,1]$

Sutton & Barto

policy $\pi(a_t \mid s_t)$ = prob. of taking action a_t in state s_t

reward $R(s_t, a_t)$ = reward for action a_t in state s_t

notation

$$p(s' \mid s, a) = P(S_{t+1} = s' \mid S_t = s, A_t = a)$$

R.V.s. S_{t+1}, S_t, A_t

Ex: $f_1(s)$ = expected (avg) 1-step reward starting from s

$$= \sum_a \pi(a \mid s) R(s,a)$$

$$= E_{a \sim \pi(\cdot \mid s)} [R(s,a)]$$

sampled from

$$= E[R(s,a) \mid s]$$

$$= E[R(S_t, A_t) \mid S_t = s]$$

Value functions

discount factor policy initial state

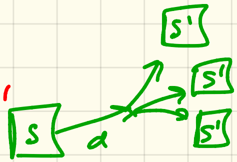
$$V_\pi(s) = E\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \,\middle|\, \pi,\ s_0 = s\right]$$

$$Q_\pi(s,a) = E\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \,\middle|\, \pi,\ s_0 = s,\ a_0 = a\right]$$

Relationships:

$$s_t \xrightarrow[a_t]{r_{t+1}} s_{t+1}$$

$$Q(s_t, a_t) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots$$



$$\downarrow$$
$$R(s,a)$$

$$V_\pi(s) = E_{a \sim \pi(\cdot|s)}\left[Q_\pi(s,a)\right]$$

$$Q_\pi(s,a) = R(s,a) + \gamma E_{s' \sim p(\cdot|s,a)}\left[V_\pi(s')\right]$$

$$Q_\pi(s,a) = \underbrace{r_1}_{R(s,a)} + \underbrace{r_2 + r_3 + r_4 + \dots}_{+ \quad V_\pi(s')}$$

Optimality:

$$V^*(s) = \max_\pi V_\pi(s)$$

$$Q^*(s,a) = \max_\pi Q_\pi(s,a)$$

$\xrightarrow{\text{agent}}$ best action maximizes $Q^*(s,a)$

$$\pi^*(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_{a'} Q^*(s,a') \\ 0 & \text{otherwise} \end{cases}$$

RL – learning from sampled action / reward / states

SARSA

Given $s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1}$

$$Q(s_t, a_t) = r_{t+1} + \gamma \, \mathbb{E}_{s_{t+1} \sim p(\cdot \mid s_t, a_t),\; a_{t+1} \sim \pi(\cdot \mid s_{t+1})} \left[Q(s_{t+1}, a_{t+1}) \right]$$

assume s_{t+1}, a_{t+1} are sampled as above

$$Q^+(s_t, a_t) \approx r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) \quad \leftarrow \text{correct on average}$$

$$Q(s_t, a_t) \leftarrow (1-\alpha) Q(s_t, a_t) + \alpha Q^+(s_t, a_t)$$

$$\leftarrow Q(s_t, a_t) + \alpha \big(\underbrace{Q^+(s_t, a_t) - Q(s_t, a_t)}_{\delta_t} \big)$$

α = learning rate $\in [0, 1]$

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \big(\underbrace{\overbrace{r_{t+1} + \gamma Q(s_{t+1}, a_{t+1})}^{\text{target}} - Q(s_t, a_t)}_{\delta_t} \big)$$

pure exploitation

how do we sample a_{t+1}?

$$a_{t+1} = \arg\max_a Q(s_{t+1}, a)$$ ← greedy policy

this means that SARSA is "on-policy"

↑ a is determined by current Q

practice: use $\underline{\varepsilon\text{-greedy policy}}$:

$$a_{t+1} = \begin{cases} \text{random } a & \text{with prob. } \varepsilon \\ \arg\max_a Q(s_{t+1}, a) & \text{otherwise} \end{cases}$$

ε = prob of exploring $\in [0,1]$

thm: SARSA converges if all (s,a) pairs are visited infinitely often and policy converges to greedy.

For example, anneal $\varepsilon = 1/t$

Q-learning

Given $s_t, a_t, r_{t+1}, s_{t+1}$

$$Q(s_t, a_t) = r_{t+1} + \gamma \, E_{s_{t+1} \sim p(\cdot \mid s_t, a_t),\ a_{t+1} \sim \pi(\cdot \mid s_{t+1})} \left[Q(s_{t+1}, a_{t+1}) \right]$$

Do not assume that a_t / a_{t+1} are from Q ← off policy

Then $$Q(s_t, a_t) = r_{t+1} + \gamma \, E_{s_{t+1}} \left[\max_{a'} Q(s_{t+1}, a') \right]$$

$$Q^+(s_t, a_t) = r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a')$$

s_{t+1} is sampled

Update $$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \Big(\underbrace{r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)}_{\delta_t} \Big)$$

In practice, choose a_{t+1} from ε-greedy policy for Q

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \Big(\underbrace{r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)}_{\nabla L\,?} \Big)$$

$$L_t(q) = \frac{1}{2}\Big(\overbrace{r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a')}^{\text{target}} - q \Big)^2$$

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) - \alpha \nabla_q L_t\big(Q(s_t, a_t)\big)$$