AML, 17 Oct 2019

Reinforcement learning

Agent chooses actions, gets a (probabilistic) reward

Determine a strategy to choose actions to maximize long term reward

Exploitation vs Exploration

Exploitation → Choosing greedily

Exploration → Learn more about system

Non-associative model

Rewards are not based on current state (i.e. only one state)

- Possibility of time varying rewards

k-bandit problem

Choose among k actions

Associative model

Markov Decision Process

Markov chain

States

Transition probability matrix

$$\begin{array}{cc} & \begin{array}{ccccc} s_1 & \cdots & s_j & \cdots & s_n \end{array} \\ s_i & \begin{bmatrix} & & & & \\ & & p_{ij} & & \\ & & & & \end{bmatrix} \end{array}$$

p_{ij} : prob of $s_i \to s_j$

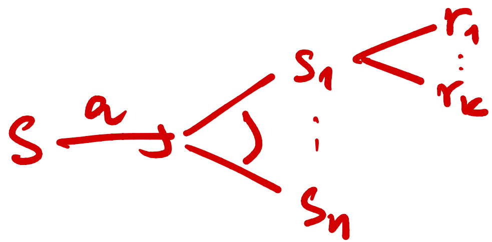
$$\sum_j p_{ij} = 1$$

Finite state, one next state distribution per state

Generalize

At s, choose a – Agent's choice

Associated probability $p(s', r \mid s, a)$



$$p(s' \mid s, a) = \sum_r p(s', r \mid s, a)$$

$$\sum_{s'} p(s' \mid s, a) = 1$$

Expected reward

$$r(s,a,s') = \sum_r r \cdot \frac{p(s', r \mid s, a)}{p(s' \mid s, a)}$$

Idealized setting for "associative" reinforcement learning

"State" information is available to agent

What does it mean to maximize long term reward?

Finite trajectory

- Play a game till it ends in win/loss/draw $\rightarrow$ "Episode"
- T moves, $t = 1, \ldots, T$

Expected return at time t

$$G_t = R_{t+1} + R_{t+2} + \cdots + R_T$$

$\downarrow$

action A_t at t generates reward R_{t+1}

$$(S_t, A_t) \longrightarrow S_{t+1}, R_{t+1}$$

Many situations - no finite episode boundary

Infinite sequence of rewards

Discounted reward

$$G_t \stackrel{\text{def}}{=} R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

$$0 \leq \gamma \leq 1$$

$$= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

$\gamma = 0$ - "myopic" $\qquad \gamma \to 1$

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

$$= R_{t+1} + \gamma \big(\underbrace{R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} \cdots}_{G_{t+1}}\big)$$

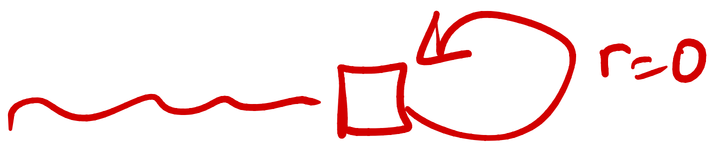
$$G_t = R_{t+1} + \gamma G_{t+1}$$

If $R = +1$ always

$$G = \frac{1}{1-\gamma}$$

To reconcile finite & infinite case

- Episodic tasks, assume that
  - terminal state have a single self loop
  - Reward is 0 for self loop



Use same discounted reward defn in both cases

Actual "value" depends on policy

$v_\pi(s)$ - value of state s wrt policy π

$$\mathbb{E}_\pi \left[G_t \mid S_t = s \right] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right]$$

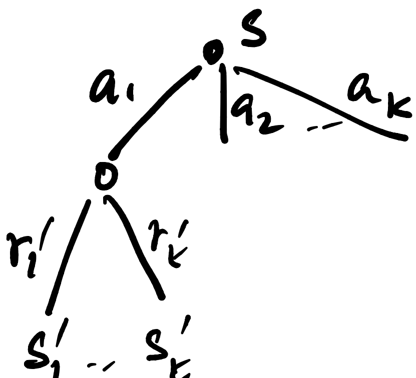
State-value function

$$q_\pi(s,a) \stackrel{\text{def}}{=} \mathbb{E}_\pi \left[G_t \mid S_t = s, A_t = a \right]$$

$$v_\pi(s) = E_\pi\left[G_t \mid S_t = s \right]$$

$$= E_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s \right]$$

$$= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma E\left[G_{t+1} \mid S_{t+1} = s' \right] \right]$$



$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s', r} p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right]$$

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_\pi(s')\right]$$

Bellman equation for $v_\pi(s)$

Find a solution to these (simultaneous) equations

Given a policy, we have Bellman equations to solve for $v_\pi(s)$

What we really want to do is to find "best" π

Best π ?

$\pi_1 \leq \pi_2$ if $\forall s.\ v_{\pi_1}(s) \leq v_{\pi_2}(s)$

There exist optimal policies

Suppose π_1, π_2 are both optimal

- $\forall s.\ v_{\pi_1}(s) = v_{\pi_2}(s)$
- $\forall s, a\ \ q_{\pi_1}(s,a) = q_{\pi_2}(s,a)$

Hypothetically, there are optimum $v_*(s)$, $q_*(s,a)$

We must have

$$
\begin{aligned}
v_*(s) &= \max_a q_*(s,a) \\
&= \max_a E\left[G_t \mid S_t = s, A_t = a\right] \\
&= \max_a E\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right] \\
&= \max_a E\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = s\right]
\end{aligned}
$$

$$
v_*(s) = \max_a \sum_{s', r} p(s', r \mid s, a)\left[r + \gamma v_*(s')\right]
$$

$$v_*(s) = \max_a \sum_{s',r} p(s',r|s,a)\left[r+\gamma v_*(s')\right]$$

vs

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r+\gamma v_\pi(s')\right]$$

Similarly

$$q_*(s,a) = \sum_{s',r} p(s',r|s,a)\left[r+\gamma \max_{a'} q_*(s',a')\right]$$