CSCI S-89C Deep Reinforcement Learning

Harvard Summer School

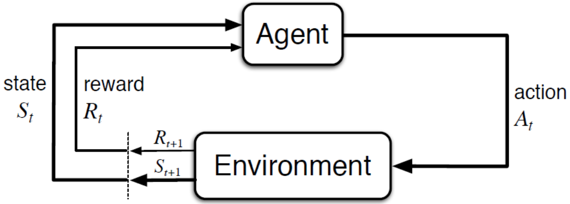
Dmitry Kurochkin

Summer 2020
Lecture 2

Contents

Contents

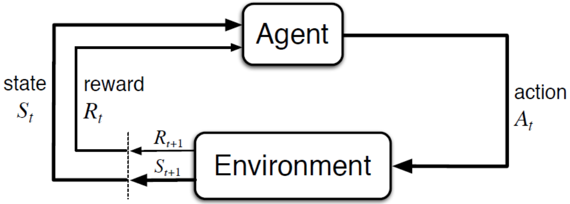
Agent–Environment Interactions



$\mathcal{S}$ = set of all possible states S_t of the environment
$\mathcal{A}(s)$ = set of all admissible actions A_t the agent can take, given $S_t = s$

Source: *Reinforcement Learning: An Introduction* by R. Sutton and A. Barto

Agent–Environment Interactions



The *trajectory* will be

$\underline{S_0, A_0}, R_1, \underline{S_1, A_1}, R_2, \underline{S_2, A_2}, R_3, \underline{S_3, A_3}, \ldots, R_{t-1}, \underline{S_{t-1}, A_{t-1}}, \ldots$

Assume Markov property, i.e. for each $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$:

$P\{S_t = s', R_t = r | S_{t-1} = s, A_{t-1} = a, R_{t-1} = r_{t-1}, S_{t-2} = s_{t-2}, \ldots\} =$
$P\{S_t = s', R_t = r | S_{t-1} = s, A_{t-1} = a\} \doteq p(s', r | s, a)$,

for any history $r_{t-1}, s_{t-2}, \ldots$

Source: *Reinforcement Learning: An Introduction* by R. Sutton and A. Barto

Agent–Environment Interactions

$p(s', r|s, a) \doteq P\{S_t = s', R_t = r | S_{t-1} = s, A_{t-1} = a\}$
completely defines the dynamics of the Markov decision processes

We notice that
$\sum_{s', r} p(s', r|s, a) = 1$
$\sum_r p(s', r|s, a) = P\{S_t = s' | S_{t-1} = s, A_{t-1} = a\} \doteq p(s'|s, a)$

Agent–Environment Interactions

$p(s', r|s, a) \doteq P\{S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a\}$
completely defines the dynamics of the Markov decision processes

We notice that
$\sum_{s',r} p(s', r|s, a) = 1$
$\sum_r p(s', r|s, a) = P\{S_t = s'|S_{t-1} = s, A_{t-1} = a\} \doteq p(s'|s, a)$

The expected reward depends on the current state s and action a only:
$E\left[R_t|S_{t-1} = s, A_{t-1} = a, R_{t-1} = r_{t-1}, S_{t-2} = s_{t-2}, \ldots\right] =$
$E\left[R_t|S_{t-1} = s, A_{t-1} = a\right] = \sum_r r \sum_{s'} p(s', r|s, a) \doteq r(s, a)$

Contents

Define Goal

$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$,
where $\gamma \in [0, 1]$ is *discount* rate

Contents

Define Policy

$$\pi(a|s) \doteq P\{A_t = a|S_t = s\}$$

Contents

State-value function $v_\pi(s)$ for policy $\pi(a|s)$

$$
\begin{aligned}
v_\pi(s) &\doteq E_\pi\left[G_t | S_t = s\right] \\
&= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s\right]
\end{aligned}
$$

Action-value function $q_\pi(s,a)$ for policy $\pi(a|s)$

$$
\begin{aligned}
q_\pi(s,a) &\doteq E_\pi\left[G_t | S_t = s, A_t = a\right] \\
&= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s, A_t = a\right]
\end{aligned}
$$

Contents

Optimal state-value function $v_*(s)$

Recall that the state-value function for policy $\pi(a|s)$ is defined as:

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

Optimal state-value function:

$$\begin{aligned} v_*(s) &\doteq \max_\pi v_\pi(s) \\ &= \max_\pi E_\pi\left[G_t | S_t = s\right] \end{aligned}$$

Existence of $\pi_*(a|s)$

Theorem

For any MDP

1. there is a policy, denoted by $\pi_*(a|s)$, that is at least as good as any other policies, i.e. it maximizes $v_{\pi_*}(s)$ for all $s \in \mathcal{S}$ simultaneously
2. optimal policy does not have to be unique, i.e. non-equal policies $\pi_{*,1}(a|s)$ and $\pi_{*,1}(a|s)$ may result in the same state-value: $v_{\pi_{*,1}}(s) = v_{\pi_{*,2}}(s)$ for all $s \in \mathcal{S}$
3. there exists a deterministic optimal policy

Optimal action-value function $q_*(s, a)$

Optimal action-value function:

$$
\begin{aligned}
q_*(s, a) &\doteq \max_\pi q_\pi(s, a) \\
&= \max_\pi E_\pi \left[G_t | S_t = s, A_t = a\right] \\
&= \max_\pi E_\pi \left[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a\right] \\
&= \max_\pi \Big[E_\pi \left[R_{t+1} | S_t = s, A_t = a\right] + \gamma E_\pi \left[G_{t+1} | S_t = s, A_t = a\right] \Big] \\
&= \max_\pi \Big[E \left[R_{t+1} | S_t = s, A_t = a\right] \\
&\qquad + \gamma \sum_{s'} p(s' | s, a) E_\pi \left[G_{t+1} \middle| S_{t+1} = s', S_t = s, A_t = a\right] \Big] \\
&= E \left[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a\right]
\end{aligned}
$$

Optimal action-value function $q_*(s, a)$

Corollary

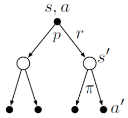
For any MDP

- any optimal policy, i.e. the policy that maximizes $v_\pi(s)$, achieves the optimal action-value $q_*(s, a)$

Contents

Bellman equation for $v_\pi(s)$

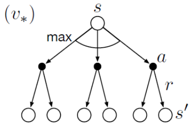
$$
\begin{aligned}
v_\pi(s) &\doteq E_\pi\left[G_t | S_t = s\right] \\
&= E_\pi\left[R_{t+1} + \gamma G_{t+1} | S_t = s\right] \\
&= \sum_a \pi(a|s) E_\pi\left[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a\right] \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)\left[r + \gamma E_\pi\left[G_{t+1} \middle| S_{t+1} = s'\right]\right] \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)\left[r + \gamma v_\pi(s')\right]
\end{aligned}
$$



Source: *Reinforcement Learning: An Introduction* by R. Sutton and A. Barto

Bellman optimality equation for $v_*(s)$

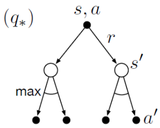
$$
\begin{aligned}
v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a) \\
&= \max_a E_{\pi_*}\left[G_t | S_t = s, A_t = a\right] \\
&= \max_a E_{\pi_*}\left[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a\right] \\
&= \max_a E_{\pi_*}\left[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a\right] \\
&= \max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_*(s')\right]
\end{aligned}
$$



Source: *Reinforcement Learning: An Introduction* by R. Sutton and A. Barto

Bellman optimality equation for $q_*(s, a)$

$$
\begin{aligned}
q_*(s,a) &= E\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \,\middle|\, S_t = s, A_t = a\right] \\
&= \sum_{s',r} p(s', r | s, a) \left[r + \gamma \max_{a'} q_*(s', a')\right]
\end{aligned}
$$



Source: *Reinforcement Learning: An Introduction* by R. Sutton and A. Barto

Contents

Iterative Policy Evaluation

Recall the Bellman equation for $v_\pi(s)$:

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

Iterative Policy Evaluation

Recall the Bellman equation for $v_\pi(s)$:

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

Let $v_k(s)$, $k = 0, 1, 2, \ldots$ denote an estimate of $v_\pi(s)$. The fixed-point iteration can be written as follows:

$$v_{k+1}(s) \doteq \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma v_k(s')\right]$$

Iterative Policy Evaluation

Let $v_k(s)$, $k = 0, 1, 2, \dots$ denote an estimate of $v_\pi(s)$. The fixed-point iteration can be written as follows:

$$v_{k+1}(s) \doteq \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma v_k(s')\right]$$

Iterative Policy Evaluation, for estimating $V \approx v_\pi$

Input π, the policy to be evaluated
Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
  $\Delta \leftarrow 0$
  Loop for each $s \in \mathcal{S}$:
    $v \leftarrow V(s)$
    $V(s) \leftarrow \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)\left[r + \gamma V(s')\right]$
    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Source: *Reinforcement Learning: An Introduction* by R. Sutton and A. Barto