10 - Policy Iteration & Policy Gradients

知识点 & 题目

Policy Evaluation

Input: π the policy for evaluation, V^π value function, and MDP $M = \langle S, s_0, A, P_a(s' \mid s), r(s,a,s') \rangle$
Output: Value function V^π

Repeat
  $\Delta \leftarrow 0$
  For each $s \in S$
    $\underbrace{V'^\pi(s) \leftarrow \sum_{s' \in S} P_{\pi(s)}(s' \mid s)\,[r(s,a,s') + \gamma\, V^\pi(s')]}_{\text{Policy evaluation equation}}$
    $\Delta \leftarrow \max(\Delta, |V'^\pi(s) - V^\pi(s)|)$
  $V^\pi \leftarrow V'^\pi$
Until $\Delta \leq \theta$

Policy Improvement

$$Q^\pi(s,a) = \sum_{s' \in S} P_a(s' \mid s)\,[r + \gamma V^\pi(s')]$$

$$\text{if } Q^\pi(s,a) > Q^\pi(s, \pi(s))$$

$$\pi(s) \leftarrow a$$

Policy Iteration

Input: MDP $M = \langle S, s_0, A, P_a(s' \mid s), r(s,a,s') \rangle$
Output: Policy π

Set V^π to arbitrary value function; e.g., $V^\pi(s) = 0$ for all s.

Set π to arbitrary policy; e.g. $\pi(s) = a$ for all s, where $a \in A$ is an arbitrary action.

Repeat $O(|A|^{|S|})$ $O(|S|^3)$

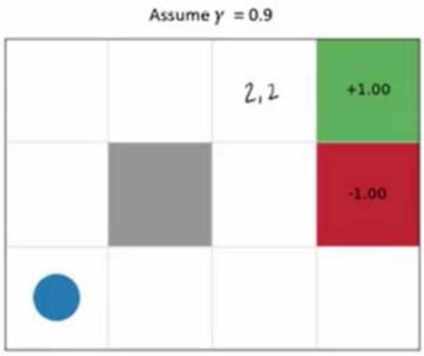
Compute $V^\pi(s)$ for all s using policy evaluation

For each $s \in S$ $O(|S|^2 \cdot |A|)$

$\pi(s) \leftarrow \text{argmax}_{a \in A(s)} Q^\pi(s,a)$

Until π does not change

POLICY ITERATION: EXAMPLE



$$V^\pi(s) = \sum_{s' \in S} P_{\pi(s)}(s' \mid s)\, [r(s,a,s') + \gamma\, V^\pi(s')]$$

$\pi(2,2) = \text{up}$

$V^\pi(3,2) = 1$

$V^\pi(2,2) = 0.8[0 + 0.9 \cdot 0] + 0.1[0 + 0.9 \cdot 0] \; 0.1[0 + 0.9 \cdot 1] = 0.09$

$Q^\pi(2,2,\text{Right}) = 0.8[0 + 0.9 \cdot 1] + 0.1[0 + 0.9 \cdot 0] + 0.1[0 + 0.9 \cdot 0] = 0.72$

$\pi(2,2) \leftarrow \text{Right}$

Policy Gradients

Algorithm – REINFORCE

Input: A differentiable policy $\pi_\theta(s,a)$, an MDP $M = \langle S, s_0, A, P_a(s' \mid s), r(s,a,s') \rangle$
Output: Policy $\pi_\theta(s,a)$

Repeat

Generate episode $(s_0, a_0, r_1, \ldots s_{T-1}, a_{T-1}, r_T)$ by following π_θ

For each (s_t, a_t) in the episode

$G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} r_k$

$\theta \leftarrow \theta + \alpha \gamma^t G \, \nabla \ln \pi_\theta(s,a)$

Until some time limit or until π_θ converges

- Continuous action space
- G: estimate of Q(s,a)
  - Instability
- Generally converge to local optima

Q Actor Critic

Input: An MDP $M = \langle S, s_0, A, P_a(s' \mid s), r(s, a, s')\rangle$

Input: A differentiable actor policy $\pi_\theta(s, a)$

Input: A differentiable critic Q-function $Q(s, a)$

Output: Policy $\pi_\theta(s, a)$

Initialise actor π parameters θ and critic parameters w arbitrarily

Repeat (for each episode)
    $s \leftarrow$ the first state in episode e
    Select action $a \sim \pi_\theta(s)$
    Repeat (for each step in episode e)
        Execute action a in state s
        Observe reward r and new state s'
        Select action $a' \sim \pi_\theta(s')$
        $\delta \leftarrow r + \gamma \cdot Q_w(s', a') - Q_w(s', a')$
        $w \leftarrow w + \alpha_w \cdot \delta \cdot \nabla Q_w(s, a)$
        $\theta \leftarrow \theta + \alpha_\theta \cdot \delta \cdot \nabla \ln \pi_\theta(s, a)$
        $s \leftarrow s'; a \leftarrow a'$

- Replace G with TD estimate -> more stable -> converge
- Critic: feedback on actions

题目