Chapter 4

- Bellman optimality equations

(4.1) $$v_*(s) = \max_a E\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right]$$

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

(4.2) $$q_*(s,a) = E\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a\right]$$

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

- Policy evaluation

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

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

- Iterative policy evaluation – the sequence $\{v_k\}$ converges to v_π as $k \to \infty$

Chapter 4

- Policy improvement

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

- Policy π' is as good as or better than π if

(4.7) $q_\pi(s, \pi'(s)) \geq v_\pi(s)$

(4.8) $v_{\pi'}(s) \geq v_\pi(s)$

- Greedy policy

(4.9) $\pi'(s) = \underset{a}{\operatorname{argmax}}\; q_\pi(s,a)$

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

- Suppose π' is as good as but not better than π

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

$\therefore$ $v_{\pi'}$ must be v_*, both π and π' are optimal

- $q_\pi(s, \pi'(s)) = \sum_a \pi'(a|s)\, q_\pi(s,a)$

for stochastic policies

Chapter 4

Policy iteration

Once a policy, π, has been improved using v_π to yield a better policy, π', we can then compute $v_{\pi'}$ and improve it again to yield an even better π''.

$$\pi_0 \xrightarrow{E} v_{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} v_{\pi_1} \xrightarrow{I} \cdots \xrightarrow{I} \pi_* \xrightarrow{E} v_*$$

evaluate improve

This method can be used to find an optimal policy.

Value iteration

$$v_{k+1}(s) = \max_a \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma v_k(s')\right] \tag{4.10}$$

↳ combines policy evaluation and policy improvement

Generalized policy evaluation (GPI)

– the general idea of letting policy evaluation and policy improvement processes interact independently of the details of the two processes.

Exercise 4.1

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

$$q_\pi(11, \text{down}) = -1 + \gamma v_\pi(11) = -1 - 14 = -15$$

$$q_\pi(7, \text{down}) = -1 + \gamma v_\pi(11) = -1 - 14 = -15$$

Exercise 4.3

(4.3) $$v_\pi(s) = E_\pi\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s\right]$$

$$q_\pi(s,a) = E_\pi\left[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a\right]$$

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

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

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

$$q_{k+1}(s,a) = \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma \sum_{a'} \pi(a'|s')\, q_k(s',a')\right]$$

- Exercise 4.4

For example, if S_{13} always transitions to S_{15} and S_{15} always transitions to S_{13}.

Possible solutions:

- Identify the "trap" states and don't allow any actions that could transition into them.

- Randomly transition out of the "trap" states and into an alternate state a small percentage of the time.

- Detect if the value is not converging by checking if the difference $|V_k - V_{k+1}|$ is getting smaller on each iteration or staying the same.

- Detect if $|V_k - V_{k+1}|$ is equal to the reward on each iteration.

- Disallow transitioning into a state once it has been identified as a "trap" state.

Chapter 4

Exercise 4.6

$$\pi_*(s) = \underset{a}{\text{argmax}} \sum_{s',r} p(s',r|s,a)\left[r + \gamma V_*(s')\right]$$

$$\pi_*(s) = \underset{a}{\text{argmax}}\ q_*(s,a)$$

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

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

$$\pi'(s) = \underset{a}{\text{argmax}} \sum_{s',r} p(s',r|s,a)\left[r + \gamma V_\pi(s')\right]$$

$$\pi'(s) = \underset{a}{\text{argmax}}\ q_\pi(s,a)$$

$$V_{k+1}(s) = \sum_{s',r} p(s',r|s,\pi(s))\left[r + \gamma V_k(s')\right]$$

$$q_{k+1}(s,a) = \sum_{s',r} p(s',r|s,a)\left[r + \gamma q_k(s',\pi(s'))\right]$$

(cont.)

1. Initialization

For each $s \in S$, $a \in A(s)$:

$Q(s,a) \leftarrow$ arbitrary

$\pi(s) \leftarrow$ arbitrary

2. Policy Evaluation

Repeat

$\Delta \leftarrow 0$

For each $s \in S$, $a \in A(s)$:

$q \leftarrow Q(s,a)$

$$Q(s,a) \leftarrow \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma Q(s', \pi(s'))\right]$$

$\Delta \leftarrow \max(\Delta, |q - Q(s,a)|)$

until $\Delta < \theta$

3. Policy Improvement

policy-stable $\leftarrow$ true

For each $s \in S$:

$a \leftarrow \pi(s)$

$$\pi(s) \leftarrow \underset{a}{\operatorname{argmax}} \overbrace{\sum_{s',r} p(s',r \mid s,a)\left[r + \gamma Q(s', \pi(s'))\right]}^{Q(s,a)}$$

If $a \neq \pi(s)$ then policy-stable $\leftarrow$ false

If policy-stable then return Q and π

else go to 2.

Exercise 4.7

A stochastic policy $\pi(a|s)$ would be used instead of a deterministic policy $\pi(s)$.

For each action a where $a \neq \pi(s)$, the stochastic policy would be:

$$\pi(a|s) = \frac{\varepsilon}{|A(s)|}$$

For action a where $a = \pi(s)$, the stochastic policy would be:

$$\pi(a|s) = 1 - \frac{\varepsilon(|A(s)|-1)}{|A(s)|}$$

$$= 1 - \frac{\varepsilon|A(s)|}{|A(s)|} + \frac{\varepsilon}{|A(s)|}$$

$$= 1 - \varepsilon + \frac{\varepsilon}{|A(s)|}$$

Step 3 would have to compute $\pi(a|s)$ in addition to computing $\pi(s)$

Step 2 would use $\pi(a|s)$ instead of $\pi(s)$

Step 1 would initialize $\pi(a|s)$ to satisfy ε-soft

Exercise 4.10

$$v_{k+1}(s) = \max_a E\left[R_{t+1} + \gamma v_k(S_{t+1}) \,\middle|\, S_t = s, A_t = a\right]$$

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

$$q_{k+1}(s,a) = E\left[R_{t+1} + \gamma \max_{a'} q_k(S_{t+1},a') \,\middle|\, S_t = s, A_t = a\right]$$

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