Margarita Prikhodko

1. a) $q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = a] =$

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

$q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = a] = E_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a] =$

$= E_\pi[R_{t+1} + \gamma \sum_{s',a'} q_\pi(s',a') \mid S_t = s, A_t = a]$

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

$q_{k+1}(s,a) = E_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a]$

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

b) $v_{k+1}(s) = E_\pi[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s] =$

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

$= T_\pi(v_k)$

$q_{k+1} = T_\pi^q(q_k)$

Show T_π^q is γ-contracting
Where $0 < \gamma < 1$ is discounting factor.

Let's consider 2 state value functions q_1 and q_2

q_1 and q_2 2 state-action value function represented as n m dim. vectors. We have to show:

$$\| T_\pi^q(q_1) - T_\pi^q(q_2) \| \leq \gamma \| q_1 - q_2 \|$$

$\|X\|$ is Euclidean norm L_2 not of X

$$T_\pi^q(q) = R_q^\pi + R_q^\pi \cdot q$$

Difference between q_1, q_2 output!

$$T_\pi^q(q_1) - T_\pi^q(q_2) =$$
$$= (R_q^\pi + P_q^\pi \cdot q_1) - (R_q^\pi + P_q^\pi \cdot q_2)$$
$$= P_q^\pi \cdot q_1 - P_q^\pi q_2$$

2 output operators!

$P_q^\pi \cdot q_1 \sim$

$$\| T_q^{\pi}(f_1) - T_q^{\pi}(f_2) \| =$$

$$= \| P_q^{\pi} \cdot f_1 - P_q^{\pi} \cdot f_2 \|$$

$$\| P_q^{\pi} f \cdot f_1 - P_q^{\pi} \cdot f_2 \|$$

Matrix normalization

$$\| X \| = \max (\| X \cdot v \| / \| v \|)$$

From above let shows

that $\| P_q^{\pi} \| \leq \gamma$

$$\| P_q^{\pi} \cdot f_1 - P_q^{\pi} f_2 \| \leq \| P_q^{\pi} \cdot (f_1 f_2) \|$$

$$\| P_q^{\pi} \cdot (f_1 - f_2) \| \leq \| P_q^{\pi} \| \cdot \| f_1 - f_2 \|$$

When $\| P_q^{\pi} \| \leq \gamma$

$$\| P_q^{\pi} (f_1 - f_2) \| \leq \gamma \| f_1 - f_2 \|$$

$$\Rightarrow$$

$$\| T_{\pi}^{q}(f_1) - T_{\pi}^{q}(f_2) \| \leq \gamma \| f_1 - f_2 \|$$

T_{π}^{q} is contracting

R_π^q - Reward for Taking ~~acting~~ action under Policy π

P_q^π State transition probabilities under policy

The property P_q^π of the q operator satisfies the Bellman equation for Q-function

$$Q_\pi(s,a) = R_q^\pi(s,a) + \gamma \sum P_q^\pi(s'|s,a) \cdot \max_{a'} Q_\pi(s'a') \quad s'a'$$

Equation is Q-value for state-action pair (s,a) depends on the immediate reward and discount factor γ.

The contraction property arises from Bellman equation Q-learning iteration

2. Exercise 4.5

1. Initialize

$Q(s,a) \in \mathbb{R}$ and

$\pi(s) \in A(s)$ arbitrarly

for $s \in S$, $a \in A$

Evaluation

Loop:

$\Delta \leftarrow 0$

Loop for each $s \in S$ and $a \in A$:

$q = Q(s,a)$

$Q(s,a) \leftarrow \sum_{s',r} p(s',r|s,a) \cdot$

$\cdot [r + \gamma \sum \pi(a'|s') Q(s',a')]$

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

while $\Delta < \varepsilon$ (ε- is error small pos. number)

Improvement

policy-stable $\leftarrow$ true

For each $s \in S$ and $a \in A$

old-action $\leftarrow \pi(s)$

$\pi(s) \leftarrow \arg\max_a Q(s,a)$

If old action $\neq$ $\pi(s)$

Then policy-stable $\leftarrow$ false

If policy-stable $Q \approx v_* \approx \pi_*$

If policy stable
then stop and return
$Q \approx q_*$ and $\pi \approx \pi_*$
else go to 2.