Assignment 5

Alec Bertelli

(1)

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

We need equations that are analogous to 4.5 and 4.4:

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

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

$$q_\pi(s,a) = \mathbb{E}_\pi\left[R_{t+1} + \gamma \sum_{a'} \pi(a'|s') q_\pi(s',a') \mid S_t = s, A_t = a\right]$$

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

For an equation that is analogous to 4.5:

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

(B) $T_q^\pi(q) = R_q^\pi + \gamma P_q^\pi q$ where $\|P_q^\pi\|_\infty \leq \gamma$

When solving for $T_q^\pi(q_1)$ and $T_q^\pi(q_2)$ we are multiplying each by the transition matrix P_q^π

If $\|P_q^\pi\|_\infty \leq \gamma$, this represents that the max discount factor to any state-action pair is γ. So this can be said as the impact of future values on state-action value functions $T_q^\pi(q_1)$ and $T_q^\pi(q_2)$ is <u>bounded</u> by γ

The contraction of T_q^π is attained by limiting the maximum row sum norm of P_q^π to γ, reducing the impact of future values on state action value functions. This <u>guarantees</u> ~~that~~ T_q^π converges to a single fixed point

(2)

We can switch the policy to use Q where the following is modeled:

$\pi_0 \rightarrow Q^{\pi_0} \rightarrow \pi_1$

where policy evaluation step uses a sequence of Q functions until the algorithm converges

We can create the following algorithm:

Init state action value function randomly for all s, a in $Q(s,a)$
Randomize the policy and set convergence threshold

while not converged:
  For each state s do:
    for each action a do:
      calculate $Q'(s,a) = \sum_{s',r} P(s',r \mid s,a)[r + \gamma \max_{a'} Q(s',a')]$

      set change_amount to max of $[|Q'(s,a) - Q(s,a)|]$
      set $Q(s,a)$ to $Q'(s,a)$

  if change_amount < convergence_threshold
    exit() the loop

Policy improvement:
  For each state s, do:
    Select action a that maximizes $Q(s,a)$: $a^* = \arg\max_a Q(s,a)$
    Update the policy for state s with Q: $\pi(s) = a^*$

Policy iteration guarantees convergence to optimal policy π and the optimal action-value function Q^* as long as the MDP is satisfied