Ex 4.3

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

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

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

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

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

Ex 4.5

```
1. Initialization.
    $q(s,a) \in R$ and $\pi(s) \in A(s)$ arbitrarily for all $s \in S$

2 Policy Evaluation
  Loop:
    $\Delta \leftarrow 0$
    Loop for each $s \in S$
      $q \leftarrow q(s, \pi(s))$
      $q(s,\pi(s)) \leftarrow \sum_{s',r} p(s',r|s,a)[r + r[\sum_{a'} \pi(a'|s') \cdot q_\pi(s',a')]]$
      $\Delta \leftarrow \max(\Delta, |q - q(s,\pi(s))|)$

3 Policy Improvement
  Policy-stable $\leftarrow$ true
  For each $s \in S$:
    old-action $\leftarrow \pi(s)$
    $\pi(s) \leftarrow \max(\sum_{s',r} p(s',r|s,a)[r + \gamma[\sum_{a'} \pi(a'|s') \cdot q_\pi(s',a')]]$
    If old-action $\neq \pi(s)$, then policy-stable $\leftarrow$ false
If policy-stable, then stop and return $q \approx q_*$ and $\pi \approx \pi_*$; else go
                                                                        to 2.
```

Question 2

Part 3 : The line in the DP method reach nearly 1.0 when state is abaut 25, than keep steady in the rest of states. And it is very smooth. However, the line in the Monte Carlo method which is in episode 8000 increases very slowly compare to the line in DP method, and it is not smooth. Because the Monte Carle method of learning is depending on experience, it trys the random policy. In gambler's problem, all possibilities at states and action are exposed to agent, and the possibilities are finite. So DP is more suitable