Q4.1

π is a equiprobable random policy

$$q_\pi(11, \text{down}) = -1 + v_\pi(T) = -1 + 0 = -1$$

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

Q4.2

$$v_\pi(15) = -1 + 0.25(-20 - 22 - 14 - v_\pi(15)) = -15 + 0.25 v_\pi(15)$$

$$v_\pi(13) = -20$$

Changing the dynamics will not result the recalculation of the ~~value~~ whole game; the set s' of $s=15$ is exactly as the one of $s=13$. They must share the same state value as -20.

Q4.3

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

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

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

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

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

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

Q4.4

1. One problem is the argmax. Argmax ~~tries last~~ breaks ties arbitrarily, this means that the same function can give rise to different policies

The way to solve this is to change the algorithm to take the whole set of maximal actions on each step and see if this set is stable & see if the policy with respect to choosing action from this set.

Q.4.5

1. Initialization
   $Q(s,a) \in \mathbb{R}$ and $\pi(s) \in A(s)$ arbitrarily for all $s \in S, a \in A$
2. Policy Evaluation
   Loop:
   $\Delta \leftarrow 0$
   Loop for each $s \in S$ and $a \in A$:
   $q = Q(s,a)$
   $Q(s,a) \leftarrow \sum_{s',r} p(s', r | s, a)[r + \gamma \sum_{a'} \pi(a'|s') Q(s', a')]$
   $\Delta \leftarrow \max(\Delta, |q - Q(s,a)|)$
   until $\Delta < \theta$
3. Policy Improvement
   policy-stable $\leftarrow$ true
   For each $s \in S$ & $a \in A$:
   old-action $\leftarrow \pi(s)$
   $\pi(s) \leftarrow \arg\max_a Q(s,a)$
   If old-action $\notin \{a, \}$, which is ~~set of~~ ~~queue~~ best solution for $\pi(s)$
   Then policy-stable $\leftarrow$ false
   If policy-stable, then stop & return $Q \approx q_*$ & $\pi \approx \pi_*$, else go to step 2

Q 4.6

Changes in step 3:

We will only decide policy-stable is false under the conditions that the policy does not explore

Changes in step 2:

Q should not be set above the limit of any soft-ϵ method

Changes in step 1:

π should be well defined as soft ϵ method ϵ should be given

Q.4.8

Since the coin is biased against us, we want to minimize the number of flips that we take. At 50 we can win with probability 0.4. At 51, if we bet small then we can get up to 52, but if we lose then we are still only back to 50 & we can again win with probability 0.4

Q.4.10

$$q_{k+1}(s,a) = E[R_{t+1} + \max_{a'} \gamma Q_k(s',a')]$$

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