Problem 1:

$$Q_{k+1}(s,a) \leftarrow \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} Q_k(s',a') \right]$$

Problem 4:

1. Start with an arbitrary initial approximation of V(s)
2. On each iteration, update the value function estimate:

$$V_{k+1}(s) \leftarrow \max_{a} \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s') \right]$$

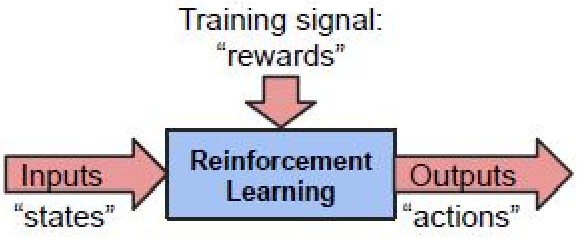
3. Stop when max value change between iterations is below a threshold.

Problem 6:

We can set the noise 0 and in that way since the agent will always end up in intended state, agent will cross the bridge because it will laurn that max reward can be received in state that has reward +10, and will know that it can not end up in state with reward -100.

Problem 10:

A. In all the experiments we our agent relly on state rewards actions. Where agent in a specific state, tries to take an action and wait for a rewards from the invirements.



B. I tried all possible combinations for epsilon and learning rate but it didn't got close because 50 iterations is not enough for the agent to find the optimal path.because the agent takes random steps to discover all possible paths it can take which means 50 iterations in not enough.

C. Epsilon-greedy used while training the agent. While using epsilon-greedy in training process it helps us to find the best policy.

D. I will set epsilon-greedy to 1 and increase the number of iteration.