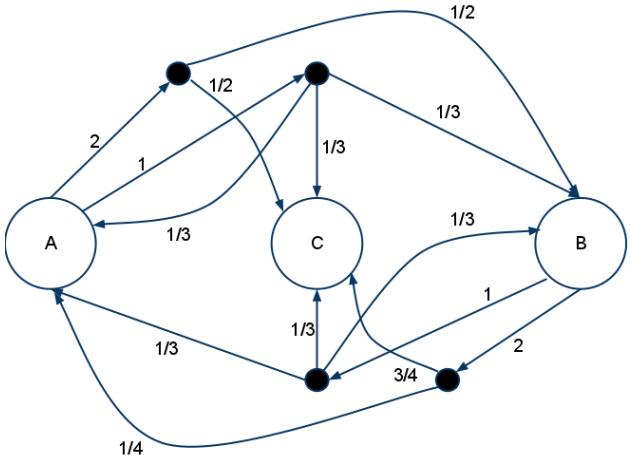
1 Policy Iteration (16 pts)

Consider the MDP below with states $X \in \{A, B, C\}$. State C is terminal. From states A and B one can take one of two actions $a \in \{1, 2\}$:

- Taking action 1 yields state A, B or C with equal probability 1/3.
- Taking action 2 from state A yields state B or C with equal probability 1/2.
- Taking action 2 from state B yields state A with probability $1/4$ and state C with probability $3/4$.

The reward function depends only on the starting state, and is $R(A) = R(B) = -1$.



1. (8pts) Assume the initial policy is $\pi(A) = 1$, $\pi(B) = 2$. Perform two iterations of policy evaluation to determine the value functions $V_2^\pi(A)$ and $V_2^\pi(B)$.
2. (8pts) Next execute a policy improvement step to determine a better policy.

2 Q-Learning (16 pts)

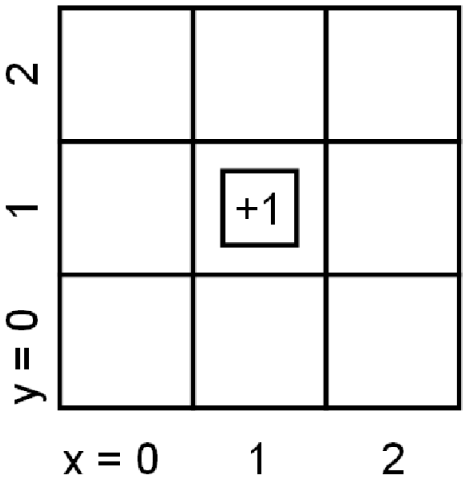
For the previous MDP, you will perform Q-learning instead. The following episodes are observed, where each trial is of the form s, a, r, where s is the current state, a is the action from that state, and r is the reward. The destination state s' is on the next line.

A, 1, -1	B, 2, -1
A, 1, -1	B, 2, -1
B, 2, -1	A, 1, -1
C	C

1. (12 pts) Perform Q-learning using both episodes.
2. (4 pts) Perform policy extraction afterwards.

3 Functional Approximation (16 pts)

For the following gridworld problem, the agent can take the actions N, S, E, W, which move the agent one square in the respective directions. There is no noise, so these actions always take the agent in the direction attempted, unless into a wall in which case the agent stays put. The boxed +1 square is the terminal state. The reward for all transitions is zero, except the transition into the terminal state, which has reward +1. Assume a discount of 0.5.



Clearly there are a number of equivalent optimal policies. Suppose we represent the values by features $f_1(x, y) = (x - 1)^2$ and $f_2(x, y) = (y - 1)^2$. Solve for the weights w_1 and w_2 that yield an optimal policy.