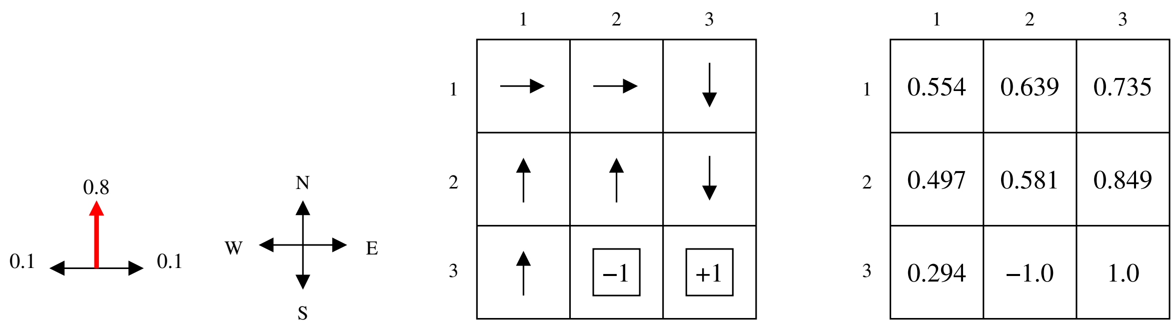
# 6.825 Reinforcement Learning
# Examples

TAs: Meg Aycinena and Emma Brunskill

## 1 Mini Grid World




(a) Transition model of 3x3 world.

(b) Optimal policy $\pi$ of 3x3 world.

(c) Utility function of optimal policy of 3x3 world.

Figure 1: A 3x3 grid world.

In the mini grid world shown in Figure 1, there are two terminal states: state (3,2) with a negative reward of $-1$, and (3,3) with a positive reward of $+1$. According to the transition model shown in Figure 1(a), an action succeeds with probability 0.8, and goes to the left or right of the intended direction with probability 0.1, respectively. The optimal policy $\pi$ is shown in Figure 1(b), and the correct utility function the optimal policy is shown in Figure 1(c).

## 2 Passive Learning

We conduct a series of three trials in this environment. We start in the start state (1,1), take actions according to the fixed policy $\pi$ given in Figure 1(b), and end once a terminal state is reached. The trials are as follows:

transitions/
trials

| Trial | $\langle$state,reward$\rangle$ series |
|---|---|
| 1 | $\langle(1,1),0\rangle \xrightarrow{E} \langle(1,2),0\rangle \xrightarrow{E} \langle(1,3),0\rangle \xrightarrow{S} \langle(2,3),0\rangle \xrightarrow{S} \langle(3,3),1\rangle$ |
| 2 | $\langle(1,1),0\rangle \xrightarrow{E} \langle(1,2),0\rangle \xrightarrow{S} \langle(2,2),0\rangle \xrightarrow{N} \langle(1,2),0\rangle \xrightarrow{E} \langle(1,3),0\rangle \xrightarrow{S} \langle(2,3),0\rangle \xrightarrow{S} \langle(3,3),1\rangle$ |
| 3 | $\langle(1,1),0\rangle \xrightarrow{S} \langle(2,1),0\rangle \xrightarrow{N} \langle(1,1),0\rangle \xrightarrow{E} \langle(1,2),0\rangle \xrightarrow{E} \langle(1,3),0\rangle \xrightarrow{S} \langle(2,3),0\rangle \xrightarrow{S} \langle(3,3),1\rangle$ |

# 3 Active Learning

## 3.1 Q-Learning

Q-learning is an alternate TD method that learns values on state-action pairs, $Q(s, a)$, instead of utilities on states. The TD update equation for Q-learning is:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(N(s,a))(R(s) + \gamma \max_{a'} Q(s', a') - Q(s,a)) \tag{4}$$

where $N(s, a)$ denotes the number of times we have been in state $s$ and taken action $a$, and $\alpha(n)$ is a function that increases and $n$ decreases. Again we use $\alpha(n) = \frac{1}{n}$.

Initialize all $Q(s, a)$ to 0. (This can also be done randomly.) Then, for each trial, we perform the update equation for each $\langle s, a, s' \rangle$ tuple. We also keep a table of $N(s, a)$ counts.

We will use the same set of trials as for the previous two examples, for simplicity. However, in general, since Q-learning is an *active* learning algorithm, each trial would have been produced using an exploration function that trades off between exploring the state-action space, and exploiting the current learned model.

Also, in the version of Q-learning presented in Russell and Norvig (page 776), a terminal state cannot have a reward. However, this is just an artifact of their particular formulation, and not inherent to Q-learning. Thus, for consistency with the previous examples, we will simply set the value of $Q(s, a)$ for a terminal state $s$ to its reward $R(s)$, for all values of $a$.

#### Trial 1

As we said above, we will learned in this trial that state $(3, 3)$ is a terminal state with reward 1. Therefore, we will set the value of the Q-function for $(3, 3$ to 1, for all $a$. And since all other utilities are zero, this is the only non-trivial update while performing updates for trial 1.

$$\begin{aligned}
Q((3,3), N) &\leftarrow 1 \\
Q((3,3), E) &\leftarrow 1 \\
Q((3,3), S) &\leftarrow 1 \\
Q((3,3), W) &\leftarrow 1
\end{aligned}$$

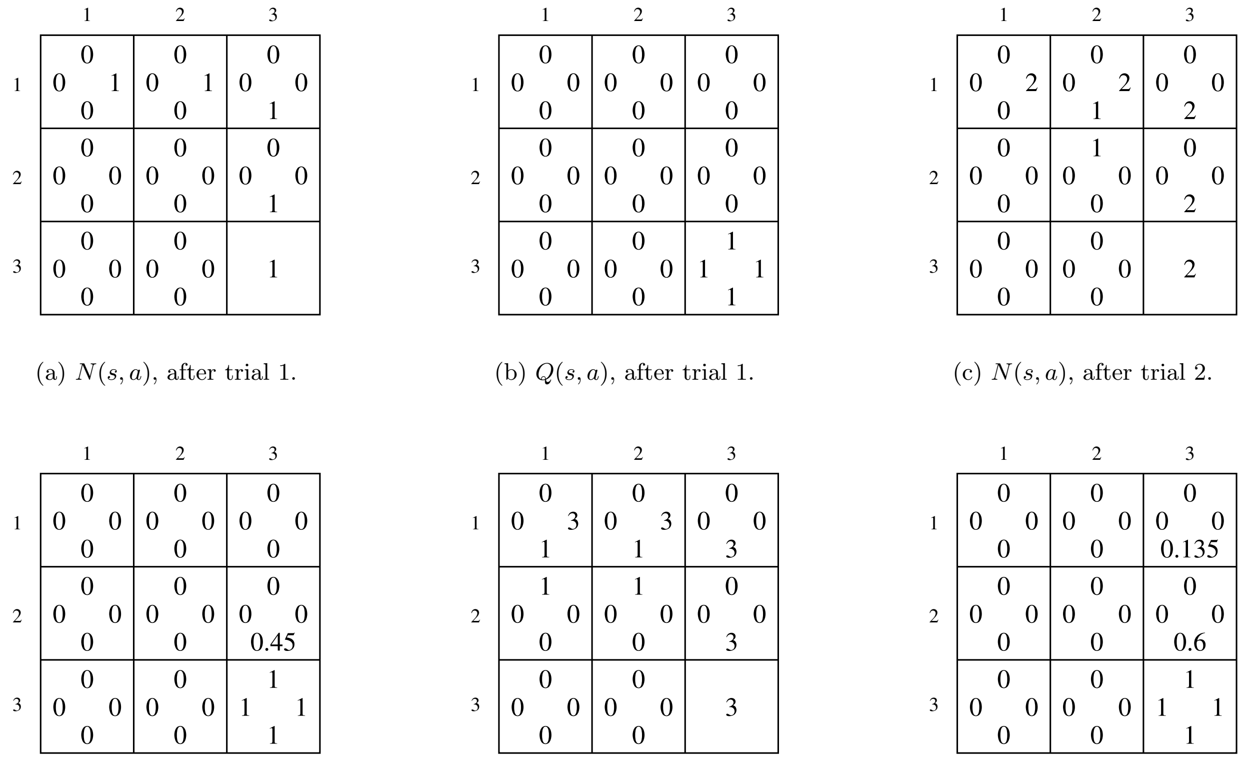
The resulting $Q(s, a)$ is shown in Figure 4(b).

#### Trial 2

The only non-zero update in this trial is for state-action pair $\langle (2, 3), S \rangle$:

$$\begin{aligned}
Q((2,3), S) &\leftarrow 0 + \frac{1}{2}(0 + 0.9(1) - 0) \\
&\leftarrow 0.45
\end{aligned}$$

The resulting $Q(s, a)$ is shown in Figure 4(d).




(a) $N(s,a)$, after trial 1.

(b) $Q(s,a)$, after trial 1.

(c) $N(s,a)$, after trial 2.

(d) $Q(s,a)$, after trial 2.

(e) $N(s,a)$, after trial 3.

(f) $Q(s,a)$, after trial 3.

Figure 4: Learned utilities, using Q-learning.

**Trial 3**

As in passive TD learning, now we have two non-zero updates during this trial; first state $(1,3)$, and then another update to state $(2,3)$:

$$
\begin{aligned}
Q((1,3),S) &\leftarrow 0 + \frac{1}{3}(0 + 0.9(0.45) - 0) \\
&\leftarrow 0.135 \\
Q((2,3),S) &\leftarrow 0.45 + \frac{1}{3}(0 + 0.9(1) - 0.45) \\
&\leftarrow 0.6
\end{aligned}
$$

The resulting $Q(s,a)$ is shown in Figure 4(f).