T-79.5102 **Autumn 2006**
Special Course in Computational Logic
Tutorial 8
Solutions

1. (a) Let us present state transitions as a graph:

$$
\begin{array}{ccccccccc}
(2,3) & \overset{\longleftarrow}{\scriptstyle 0.25} & (2,3) & \overset{\longleftarrow}{\scriptstyle 0.25} & (\mathbf{2},\mathbf{3}) & \overset{\longrightarrow}{\scriptstyle 0.25} & (3,3) & \overset{\longrightarrow}{\scriptstyle 0.25} & (3,3) \\
 & {\scriptstyle 0.50}\swarrow & \downarrow^{0.25} & & \downarrow^{0.50} & & \downarrow^{0.25} & \searrow_{0.50} & \\
 & (2,2) & (3,3) & & (2,2) & & (2,3) & (3,2) & \\
 & & & {\scriptstyle 0.25}\swarrow & \downarrow^{0.50} & \searrow_{0.25} & & & \\
 & & & (1,2) & (2,1) & (3,2) & & &
\end{array}
$$

Then we may summarise probabilities for individual states:

$$
\begin{aligned}
P(1,2) &= 0.50 \times 0.25 = 0.125 \\
P(2,1) &= 0.50 \times 0.50 = 0.25 \\
P(2,2) &= 0.25 \times 0.50 = 0.125 \\
P(2,3) &= 0.25 \times 0.25 + 0.25 \times 0.25 = 0.125 \\
P(3,2) &= 0.50 \times 0.25 + 0.25 \times 0.50 = 0.25 \\
P(3,3) &= 0.25 \times 0.25 + 0.25 \times 0.25 = 0.125
\end{aligned}
$$

The sum of probabilities is 1 (as it should).

(b) We begin by writing down a set of equations for the expected utilities u_{ij} for each state (i,j):

$$
\begin{cases}
u_{12} = -0.25 + 0.5u_{12} & (1) \\
u_{23} = -0.25 + 0.5a + 0.25u_{23} & (2) \\
u_{22} = -0.25 + 0.5u_{21} + 0.25u_{12} - 0.25 = 0 & (3) \\
u_{21} = -0.25 + a + 0.25u_{21} & (4)
\end{cases}
$$

Note in particular how the cost -0.25 of a move is incorporated in each equation. The set of equations is solved as follows.

(1) $\implies 0.5u_{12} = -0.25 \implies u_{12} = -0.5.$

(3) $\implies 0.5u_{21} = 0.5 - 0.25u_{12} = 0.625 \implies u_{21} = \frac{0.625}{0.5} = 1.25.$

(4) $\implies a = 0.75u_{21} + 0.25 = 1.1875$

(2) $\implies 0.75u_{23} = 0.5a - 0.25 \implies u_{23} = \frac{0.5a-0.25}{0.75} \approx 0.4583.$

Thus $u_{12} = -0.5$, $u_{21} = 1.25$, $u_{23} \approx 0.4583$, $a = 1.1875$ ja $2a = 2.375$.

(c) Let us calculate the expected utility u_{21} when $\leftarrow$ is the action assigned to $(2,1)$ by the policy:

$u_{12} = -0.25 + 0.50u_{12} + 0.25u_{12} + 0.25u_{12}$

$\implies u_{12} = -0.25 + u_{12}$

$\implies 0 = -0.25.$

There is no solution, i.e., the expected utility u_{21} cannot be determined. This is because $u_{21} \longrightarrow -\infty$.

2. Given the simplified (fully observable) grid environment

■		+1
S		-1

the state space of the agent is $S = \{(1,1),(2,1),(3,1),(2,2),(3,2)\}$ and the set of possible actions $A = \{\leftarrow, \uparrow, \rightarrow, \downarrow\}$.

A *policy* π is an arbitrary function from S to A. In other words, a policy attachs a unique action $a = \pi(s)$ to each state s, and the agent executes a every time it is in s. An optimal policy π^* assigns to each state s an action $a = \pi^*(s)$ that maximises the expected utility $\mathrm{EU}_s(a) = \sum_{s'} T(s,a,s')U(s')$ where $T(s,a,s')$ gives the transition probability from s to s'. Note that $\sum_{s'} T(s,a,s') = 1$ holds for each state s and action a.

(a) The *value iteration* algorithm computes iteratively the new utility values for each state s:

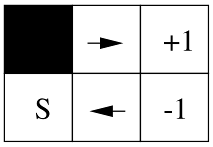
$$U_{i+1}(s) = R(s) + \max_a \textstyle\sum_{s'} T(s,a,s')U_i(s')$$

where $R(s)$ is the *reward* of the state (here 1 in $(3,2)$, -1 in $(3,1)$, and -0.2 in all other states). Such a calculation is repeated until utility values converge, i.e., $|U_{i+1}(s) - U_i(s)|$ becomes small enough for each state s. Then the action with the maximum expected utility is chosen as $\pi^*(s)$ for a particular state s.

Round $i = 0$:

State s	a	$\mathrm{EU}_s(a)$	
$(2,2)$	$\leftarrow$	$1 \cdot (-0.2) = -0.2$	
	$\uparrow$	$0.9 \cdot (-0.2) + 0.1 \cdot 1 = -0.08$	
	$\rightarrow$	$0.8 \cdot 1 + 0.2 \cdot (-0.2) = 0.76$	$\times$
	$\downarrow$	$0.9 \cdot (-0.2) + 0.1 \cdot 1 = -0.08$	
$(2,1)$	$\leftarrow$	$1 \cdot (-0.2) = -0.2$	$\times$
	$\uparrow$	$0.9 \cdot (-0.2) + 0.1 \cdot (-1) = -0.28$	
	$\rightarrow$	$0.8 \cdot (-1) + 0.2 \cdot (-0.2) = -0.84$	
	$\downarrow$	$0.9 \cdot (-0.2) + 0.1 \cdot (-1) = -0.28$	
$(1,1)$	$\leftarrow$	$1 \cdot (-0.2) = -0.2$	
	$\uparrow$	$1 \cdot (-0.2) = -0.2$	
	$\rightarrow$	$1 \cdot (-0.2) = -0.2$	
	$\downarrow$	$1 \cdot (-0.2) = -0.2$	

So, the optimal action in $(2,2)$ is $\rightarrow$ and in $(2,1)$ it is $\leftarrow$. Since all actions have the same expected utilities in $(1,1)$ the choice is free:



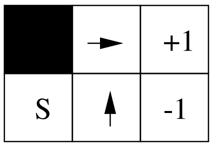
The new expected utilities are:

$$U_1(2,2) = -0.2 + 0.76 = 0.56$$
$$U_1(2,1) = -0.2 - 0.2 = -0.4$$
$$U_1(1,1) = -0.2 - 0.2 = -0.4$$

Round $i = 1$:

State s	a	$\mathrm{EU}_s(a)$	
$(2,2)$	$\leftarrow$	$0.9 \cdot 0.56 + 0.1 \cdot (-0.4) = 0.464$	
	$\uparrow$	$0.9 \cdot 0.56 + 0.1 \cdot 1 = 0.604$	
	$\rightarrow$	$0.8 \cdot 1 + 0.1 \cdot 0.56 + 0.1 \cdot (-0.4) = 0.816$	$\times$
	$\downarrow$	$0.8 \cdot (-0.4) + 0.1 \cdot 0.56 + 0.1 \cdot 1 = -0.164$	
$(2,1)$	$\leftarrow$	$0.9 \cdot (-0.4) + 0.1 \cdot 0.56 = -0.304$	
	$\uparrow$	$0.8 \cdot 0.56 + 0.1 \cdot (-1) + 0.1 \cdot (-0.4) = 0.308$	$\times$
	$\rightarrow$	$0.8 \cdot (-1) + 0.1 \cdot (-0.4) + 0.1 \cdot 0.56 = -0.784$	
	$\downarrow$	$0.9 \cdot (-0.4) + 0.1 \cdot (-1) = -0.46$	
$(1,1)$	$\leftarrow$	$1 \cdot (-0.4) = -0.4$	
	$\uparrow$	$1 \cdot (-0.4) = -0.4$	
	$\rightarrow$	$1 \cdot (-0.4) = -0.4$	
	$\downarrow$	$1 \cdot (-0.4) = -0.4$	

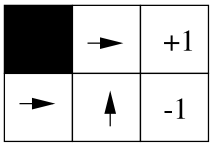
The resulting policy is



and the new utility values are

$$
\begin{aligned}
U_2(2,2) &= -0.2 + 0.816 = 0.616 \\
U_2(2,1) &= -0.2 + 0.308 = 0.108 \\
U_2(1,1) &= -0.2 - 0.4 = -0.6
\end{aligned}
$$

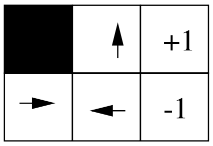
While continuing the execution of the value iteration algorithm, the optimal actions in $(2,2)$ and $(2,1)$ stay unchanged. Finally, the state $(1,1)$ gets a (unique) optimal action because the utility of $(2,1)$ becomes higher than that of $(1,1)$. Thus, the resulting policy is:



This is actually optimal but it takes still several rounds of the algorithm until the utility values stabilize.

(b) In *policy iteration* we start by creating a random policy π_0. Then, we compute the utility values of states given the policy π_i, revise the policy π_i to π_{i+1} by choosing the actions with highest expected utilities, and compute new utility values. This process is continued until the policy under construction stabilises, i.e., $\pi_{i+1} = \pi_i$.

Suppose that the following random policy π_0 is chosen:



The utilities given π_0 can be computed analytically by solving the following group of equations. In the following, u_{ij} denotes the utility of the state (i,j).

$$
\begin{aligned}
u_{11} &= 0.2u_{11} + 0.8u_{21} - 0.2 \\
u_{21} &= 0.8u_{11} + 0.1u_{21} + 0.1u_{22} - 0.2 \\
u_{22} &= 0.9u_{22} + 0.1 \cdot 1 - 0.2
\end{aligned}
$$

The solution for this set of equations is:

$$
\begin{aligned}
u_{11} &= -5.25 \\
u_{21} &= -5 \\
u_{22} &= -1
\end{aligned}
$$

Now we compute the expected utilities for different actions:

State s	a	$\mathrm{EU}_s(a)$	
$(2,2)$	$\leftarrow$	$0.9 \cdot (-1) + 0.1 \cdot (-5) = -1.4$	
	$\uparrow$	$0.9 \cdot (-1) + 0.1 \cdot 1 = -0.8$	
	$\rightarrow$	$0.8 \cdot 1 + 0.1 \cdot (-1) + 0.1 \cdot (-5) = 0.2$	$\times$
	$\downarrow$	$0.8 \cdot (-5) + 0.1 \cdot (-1) + 0.1 \cdot 1 = -4$	
$(2,1)$	$\leftarrow$	$0.8 \cdot (-5.25) + 0.1 \cdot (-5) + 0.1 \cdot (-1) = -4.8$	
	$\uparrow$	$0.8 \cdot (-1) + 0.1 \cdot (-1) + 0.1 \cdot (-5.25) = -1.425$	
	$\rightarrow$	$0.8 \cdot (-1) + 0.1 \cdot (-5) + 0.1 \cdot (-1) = -1.4$	$\times$
	$\downarrow$	$0.8 \cdot (-5) + 0.1 \cdot (-1) + 0.1 \cdot (-5.25) = -4.625$	
$(1,1)$	$\leftarrow$	$1 \cdot (-5.25) = -5.25$	
	$\uparrow$	$0.9 \cdot (-5.25) + 0.1 \cdot (-5) = -5.225$	
	$\rightarrow$	$0.8 \cdot (-5) + 0.2 \cdot (-5.25) = -5.05$	$\times$
	$\downarrow$	$0.9 \cdot (-5.25) + 0.1 \cdot (-5) = -5.225$	

The revised policy π_1 is

	$\rightarrow$	+1
$\rightarrow$	$\rightarrow$	-1

After the next round of the algorithm, the action for $(2,1)$ changes to the optimal one, i.e., $\uparrow$.