6. (10 points.) MDPs and Reinforcement Learning

Consider an autonomous robot which can either move FAST or SLOW in any time step. Moving FAST generally gives a reward of +2, while moving SLOW gives a reward of only +1. However, the robot must also take into account its internal temperature, which can be either HOT or OK. Driving SLOW tends to lower the temperature, while driving FAST tends to raise it. If the robot is HOT, there is a danger if it overheating, at which point it must stop, cool down, and make repairs. The MDP transitions and rewards are specified as follows:

s	a	s'	$T(s,a,s')$	$R(s,a,s')$
OK	SLOW	OK	1.0	+1
OK	FAST	OK	0.5	+2
OK	FAST	HOT	0.5	+2
HOT	SLOW	OK	1.0	+1
HOT	FAST	HOT	0.5	+2
HOT	FAST	OK	0.5	-10

Note that while repairs are costly, the robot is OK afterwards (the last row in the table).

(1) (5 pts): Run two rounds of value iteration in the table below, using a discount of 0.8. You may skip the greyed-out square.

s	V_0	V_1	V_2
OK	0	2	3.2
HOT	0	1	

$$V_{i+1}(s) \longleftarrow \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_i(s')]$$

$$V_1(ok) \longleftarrow \max[(1(1+\gamma*0)), (0.5(2+\gamma*0)+0.5(2+\gamma*0))] = max(1,2) = 2$$
$$V_1(hot) \longleftarrow \max[(1(1+\gamma*0)), (0.5(2+\gamma*0)+0.5(-10+\gamma*0))] = max(1,-4) = 1$$
$$V_2(ok) \longleftarrow \max[(1(1+\gamma*2)), (0.5(2+\gamma*2)+0.5(2+\gamma*1))] = max(2.6,3.2) = 3.2$$

(1) (5 pts): Run Q-learning with a discount of 0.8 and a learning rate of 0.5, using the transition samples below. Do not copy over q-values which have not changed in a given step.

Assume the agent experiences the samples:

OK, FAST, HOT, reward +2, calculate Q_1
HOT, FAST OK, reward -10, calculate Q_2
OK, SLOW, OK, reward +1, calculate Q_3

s	a	Q_0	Q_1	Q_2	Q_3
OK	SLOW	0			0.9
OK	FAST	0	1.0		
HOT	SLOW	0			
HOT	FAST	0		-4.6	

$$
\begin{aligned}
Q(s,a) &\longleftarrow Q(s,a) + 0.5[R(s,a,s') + 0.8 \max_{a'} Q(s',a') - Q(s,a)] \\
Q_1(ok, fast) &\longleftarrow 0.0 + 0.5[2.0 + 0.8 \max_{a'} Q(hot, a') - 0.0] = 1 \\
Q_2(hot, fast) &\longleftarrow 0 + 0.5[-10 + 0.8 \max_{a'} Q(ok, a') - 0] = -4.6 \\
Q_3(ok, slow) &\longleftarrow 0 + 0.5[1 + 0.8 \max_{a'} Q(ok, a') - 0] = 0.9
\end{aligned}
$$

End of Exam