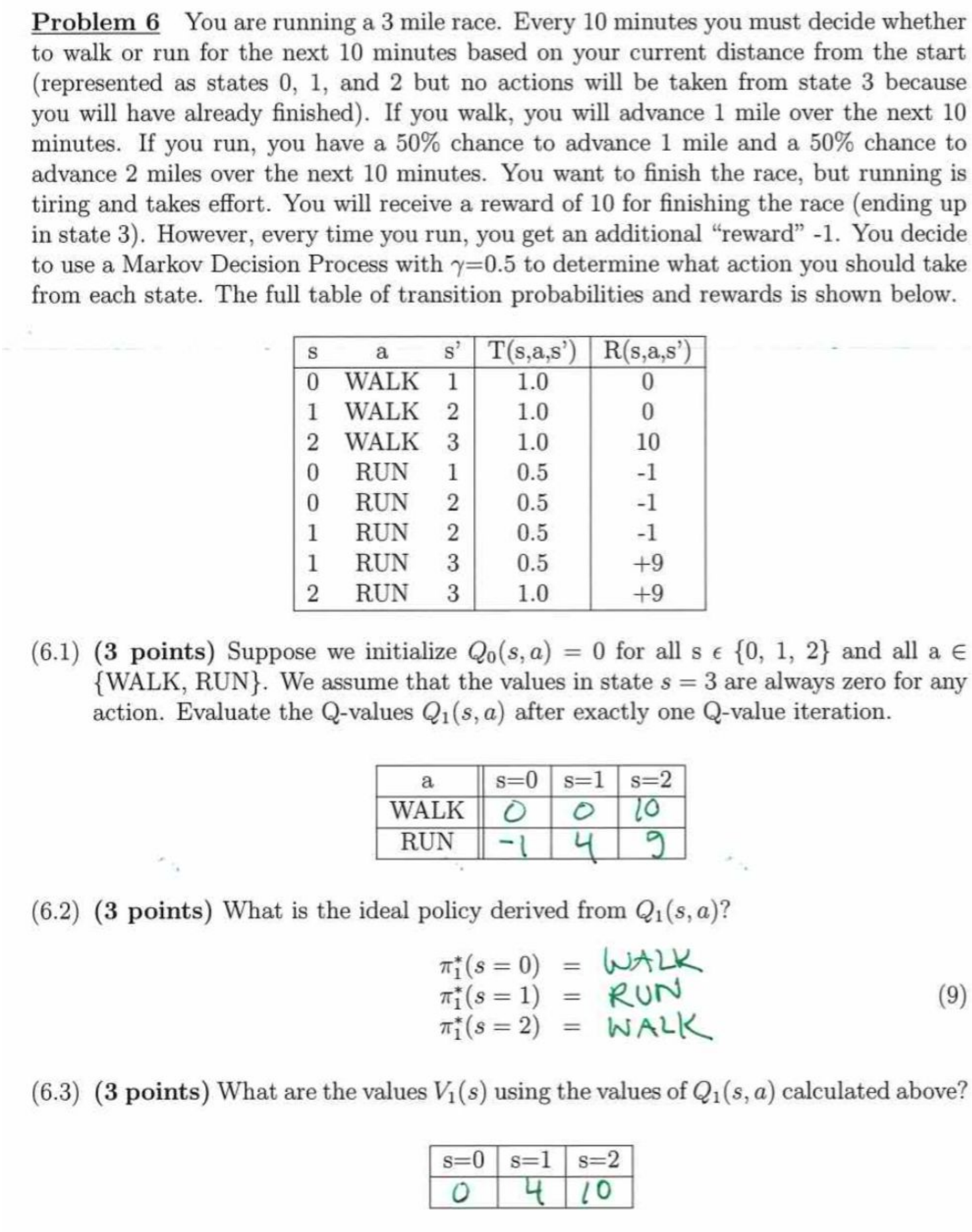
PROBLEM 8



Problem 6 You are running a 3 mile race. Every 10 minutes you must decide whether to walk or run for the next 10 minutes based on your current distance from the start (represented as states 0, 1, and 2 but no actions will be taken from state 3 because you will have already finished). If you walk, you will advance 1 mile over the next 10 minutes. If you run, you have a 50% chance to advance 1 mile and a 50% chance to advance 2 miles over the next 10 minutes. You want to finish the race, but running is tiring and takes effort. You will receive a reward of 10 for finishing the race (ending up in state 3). However, every time you run, you get an additional "reward" -1. You decide to use a Markov Decision Process with γ=0.5 to determine what action you should take from each state. The full table of transition probabilities and rewards is shown below.

s	a	s'	T(s,a,s')	R(s,a,s')
0	WALK	1	1.0	0
1	WALK	2	1.0	0
2	WALK	3	1.0	10
0	RUN	1	0.5	-1
0	RUN	2	0.5	-1
1	RUN	2	0.5	-1
1	RUN	3	0.5	+9
2	RUN	3	1.0	+9

(6.1) **(3 points)** Suppose we initialize $Q_0(s, a) = 0$ for all s ϵ {0, 1, 2} and all a $\in$ {WALK, RUN}. We assume that the values in state $s = 3$ are always zero for any action. Evaluate the Q-values $Q_1(s, a)$ after exactly one Q-value iteration.

a	s=0	s=1	s=2
WALK	0	0	10
RUN	-1	4	9

(6.2) **(3 points)** What is the ideal policy derived from $Q_1(s, a)$?

$$\begin{aligned} \pi_1^*(s=0) &= \text{WALK} \\ \pi_1^*(s=1) &= \text{RUN} \\ \pi_1^*(s=2) &= \text{WALK} \end{aligned} \tag{9}$$

(6.3) **(3 points)** What are the values $V_1(s)$ using the values of $Q_1(s, a)$ calculated above?

s=0	s=1	s=2
0	4	10

(6.4) **(3 points)** Consider now iterating Q-values one more time to obtain $Q_2(s,a)$. We are only interested in here what happens at $s = 0$. For what range of values of the discount factor $0 \leq \gamma \leq 1$ would the action derived from $Q_2(0,a)$ suggest that we RUN?

$$Q_2(0,W) = 0 + \gamma \cdot V_1(1) = \gamma \cdot 4$$

$$Q_2(0,R) = \frac{1}{2}\left[-1 + \gamma \cdot V_1(1)\right] + \frac{1}{2}\left[-1 + \gamma \cdot V_1(2)\right]$$

$$= -1 + \gamma \cdot 7$$

$Q_2(0,R) \geq Q_2(0,W)$ if $\gamma \geq \frac{1}{3}$

8) a) Solutions are fine.

b) We get the optimal policy from the largest $Q_1(s, a)$-values.

c) We can get the $V_1(s)$ values from the largest $Q(s, a)$ values.