This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the README.md for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

1.a	iteration 0	iteration 1	iteration 2
Vopt(-2)	0	0	0
Vopt(-1)	0	$\max\{\overset{16-1}{0.8\times20+0.2\times(-5)}, \overset{-1.5+14}{0.3\times(-5)+0.7\times20}\}=15$	$\overset{16}{0.8}\times20+0.2\times(-5-5))=14$
Vopt(0)	0	$\max\{\overset{-4-1}{0.8\times(-5)+0.2\times(-5)}, \overset{-1.5-3.5}{0.3\times(-5)+0.7\times(-5)})=-5$	$\max\{0.8\times(15-5)+0.2\times(26.5-5), 0.3\times26.5+0.7\times10\}=13.45$
Vopt(1)	0	$\max\{0.8\times(-5)+0.2\times100, 0.3\times100+0.7\times(-5)\}=26.5$	$0.3\times100+0.7\times(-10)=23$
Vopt(2)	0	0	0

1.b

$\pi_{opt}(-1) = -1$

$\pi_{opt}(0) = +1$

$\pi_{opt}(1) = +1$

2.b To compute Vopt for each node with only a single path.

we can add memoization to the recursion, which is equivalent

to use dynamic programming to compute the value at each node.

2.c

$$V_{opt}^{(t)}(s) \leftarrow \max_{a \in Action(s)} \sum_{s'} T(s,a,s')\left[Reward(s,a,s') + \lambda V_{opt}^{(t-1)}(s')\right]$$

$$\lambda T(s,a,s')\left[\frac{1}{\lambda} Reward(s,a,s') + V_{opt}^{(t-1)}(s')\right]$$

$T'(s,a,s') = \lambda T(s,a,s')$ for $s' \in S$

$T'(s,a,o) = 1-\lambda$

$R'(s,a,s') = \frac{1}{\lambda} Reward(s,a,s')$ for $s' \in S$

$R'(s,a,o) = 0$

4.b For small MDP, Q-learning produces less than 10% different actions.
For large MDP, Q-learning produces around 35% different actions.
what went wrong is large MDP has more unknown states that Q-learning is not able to learn accurately.
Another reason is the identity feature extractor cannot replesent the value of unknown states.

4.d FixedRLAlgorithm has rewards around 6~7
while Q-learning has rewards around 11-12.
The reason that Q-learning has higher rewards is because
Q-learning can adapt to the newThresholdMDP while
FixedRLAlgorithm is fixed and cannot adapt.