William Frank
2/20/19
Homework 5: MDPs

1. Life as a Professor
    1. Filling in the V* table:

t	V_t*(asst)	V_t*(assc)	V_t*(full)	V_t*(hl)	V_t*(dead)
0	0	0	0	0	0
1	20	60	400	10	0
2	33	119	540	13.5	0
3	43.15	151.05	589	14.725	0
4	49.5225	165.6875	606.15	15.15375	0
5	52.940875	171.836625	612.1525	15.3038125	0

    2. What was the change in max-norm of V between steps 4 and 5? What does this tell us about how close the final values are to the true values?

Th max norm is found in V_t*(full) between steps 4 and 5, and is 6.0025. As this is a relatively large value, it tells us that we are not very close to reaching the final value values.

2. Clarence the Evil Professor
    1. 4 iterations of value iteration with policy always work:

t	D+F	S+F	D+P	S+P
0	0	0	0	0
1	0	0	10	10
2	0	0	10	10
3	0	0	10	10
4	0	0	10	10

This outcome makes sense, from all states action W will bring you to a state with reward 0, so the value will never increase.

    2. 4 iterations of value iteration with policy work when dumb and sweettalk when smart:

t	D+F	S+F	D+P	S+P
0	0	0	0	0
1	0	0	10	10
2	0	2.5	10	15
3	.625	3.75	10.625	16.25
4	1.09375	4.21875	11.09375	16.71875

This is a much better policy than the previous one, all states have a value higher than their original value. This also makes sense, Alice is choosing options which will move her to reward states, increasing the value of the initial states.

3. 4 iterations of straight value iteration with no initial policy:

t	D+F	S+F	D+P	S+P
0	0	0	0	0
1	0	0	10	10
2	0	2.5	12.5	15
3	.625	3.75	13.125	16.875
4	1.09375	4.375	13.4375	17.578125

4. Policy Iteration

New Policy:

D+F:	Work
S + F:	Sweettalk
D + P:	Sweettalk
S + P:	Sweettalk

4 iterations of value iteration with this policy:

t	D+F	S+F	D+P	S+P
1	0	0	0	0
2	0	0	10	10
3	0	2.5	12.5	15
4	.625	3.75	13.125	16.875

New Policy:

D+F:	Work
S + F:	Sweettalk
D + P:	Sweettalk
S + P:	Sweettalk

The policy has converged two iterations from the initial policy. However, had I chosen Sweettalk instead of Work for D+F (they were tied), then one more iteration would be needed to converge. This does seem intuitively optimal from just looking at the initial graph, and is also the policy that value iteration “chooses”.