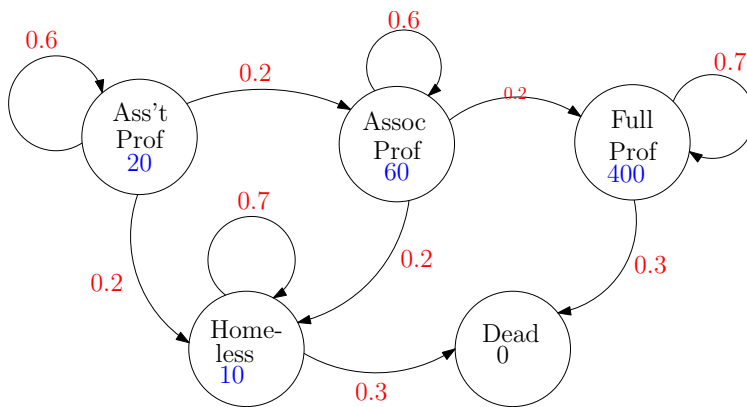


## HW4: Markov Decision Processes

## 1 Life as a Professor

Bob, our venerable AI professor, is currently riding the currents of new-professorhood. To better understand what his life will be like, he has constructed an Markov Process (like an MDP but where you don't choose actions, you just have the world act on you... you can think of this as an MDP where there is only one action to take in each state) to represent the life of a professor. He has drawn the life of a professor in the following MP, where rewards are written on the states and transition probabilities are written on the arcs. His start state is as an assistant professor.



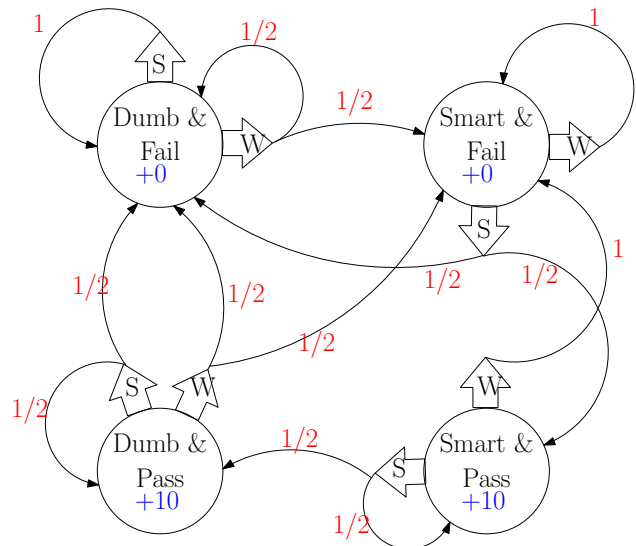
- Fill in the following table with the  $V^*$  values for each state in the MDP for  $t = 1 \dots 5$ ; I have filled in the first three values to help you out (hint: use the value iteration Bellman updates, but remove the "max<sub>a</sub>" part because you don't need to select actions). Suppose  $\gamma = 0.5$ .

$t$	$V_t^*(\text{asst})$	$V_t^*(\text{assoc})$	$V_t^*(\text{full})$	$V_t^*(\text{hl})$	$V_t^*(\text{dead})$
0	0	0	0	0	0
1	20	60	400	...	...
2	...	...	...	...	...
3	...	...	...	...	...
4	...	...	...	...	...
5	...	...	...	...	...

- 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?

## 2 Clarence the Evil Professor

Alice is also taken another class (class name withheld!) from Professor Clarence. Clarence is known to be evil among the department because he only passes students who sweettalk him. Based on what she’s learned in AI, Alice has figured out that classes with Clarence can be modeled with the following MDP:



The difference between this Figure and the one from the previous question is now Alice has actions she can take. In any state, she can either take action “S” (to try to Sweettalk Clarence) or action “W” (to Work hard). Her state is represented by a pair: they can either be smart or dumb, and she can either pass or fail the course. The transitions can be read as follows. If Alice is Dumb and Failing, and she chooses to Sweettalk, then Clarence won’t listen to her because she’s dumb, so with probability 1 she winds back up in the same state. On the other hand, had she chosen to Work, then with 50% probability she’d be successful and become Smart (but still Failing) and with 50% probability she’d have studied the wrong thing and will remain Dumb and Failing.

Please answer the following questions with respect to this MDP; always assume  $\gamma = 0.5$ :

1. Suppose that Alice’s policy is to always work. Compute the value of each state under this policy, running 4 iterations of value iterations. Does this outcome make sense?

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

2. Suppose that Alice’s policy is to work only when she’s dumb. Compute the values here for four steps. Which policy is better (and does this make sense)?

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

3. Now, suppose we do not begin with an initial policy but just run plain value iteration. Show the  $V$  values derived through value iteration for the first four time steps; fill in the table below:

<b>t</b>	D+F	S+F	D+P	S+P
0	0	0	0	0
1	0	0	10	10
2	...	...	...	...
3	...	...	...	...
4	...	...	...	...

4. **(6300 only)** Let's do policy iteration! Suppose that Alice begins with the policy from question (1): always work. You've already computed values under this policy. Use that to estimate a new policy. Write that policy down. Compute four iterations of values for that new policy and compute a third policy. Has the policy converged? Does it seem (intuitively) optimal?