Department of Electrical and Computer Engineering National University of Singapore Dr Fan Shi, Dr Armin Lederer	**Robotics and Embodied Artificial Intelligence** **CEG5306** Assignment 1: MDPs and multi-armed bandits	Autumn 2025

1. **Identifying the MDP in Gymnasium environments (~ 6 pts)**
   Gymnasium (`https://gymnasium.farama.org`) is a very popular Python package as it provides an easy-to-use API for single-agent reinforcement learning environments. Moreover, many simple environments are already implemented, such they are readily usable.
   (a) Specify the state and action spaces, the initial distribution and the reward for the Gymnasium Acrobot environment. For simplicity, ignore the termination and assume a reward of 0 is assigned to every state achieving the target height. Be mathematically accurate!
   (b) How would the reward need to be defined to specify the goal of moving the robot to the designated target height while avoiding large torques? Explain your choice!
   (c) Explain which difficulty can arise from a goal specification as in b)? Discuss only one difficulty!

2. **Finite MDPs and value functions (~ 12 pts)**
   Student A has 2 days left before he needs to submit a research paper and an assignment. A can decide every day to work either on research or on the assignment. If A decides to work on research, he has a break through, i.e., the paper is finished, with probability 0.5. If A works on the assignment, he completes the assignment with probability 0.9. Finishing the assignment is a bit of a relief for A (happiness + 1). Finishing the research paper on time makes A feel great (happiness + 10). A wants to maximize his happiness.
   (a) Draw the MDP. For this, represent each state using a circle with a number/letter in its center. Possible transitions must be illustrated via arrows with annotations indicating which action caused the transition. Note that one action can cause multiple transitions. Highlight the initial state.
   (b) Specify the state and action spaces, the initial state probabilities and a reward function describing this problem. Do not use reward functions depending on the next state and be mathematically accurate!

   When there is no deadline and infinitely many papers/assignments can be finished, we obtain the simplified MDP illustrated in Fig. 1.

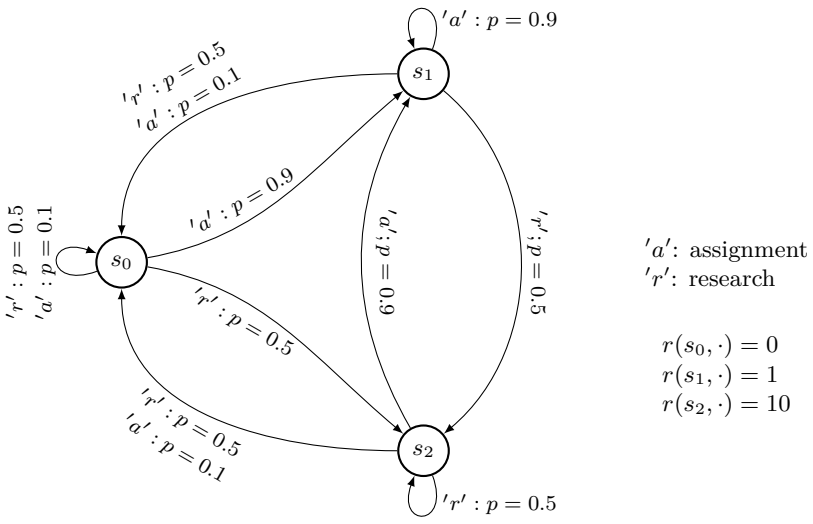

Figure 1: Illustration of the simplified MDP.

   (c) Determine the value function for the strategy of working only on assignments in this simplified scenario. Assume a discount factor $\gamma = 0.9$.

(d) Is the strategy in c) optimal? Explain your answer!

3. **Multi-armed bandits (~ 5 pts)**
Student A is needs to tune the controller for an unmanned aerial vehicle (UAV). A has already identified 5 different parameters θ_i that all seem to work well in general, but it is not clear which one is the best. Assume that the used performance is normalized to the interval $[0, 1]$ in the following.

(a) A selects parameters according to a round robin procedure, i.e., he applies θ_1, θ_2, ..., θ_5, then starts again with θ_1. He uses this approach until each parameter has been used 100 times. Derive a bound for the expected cumulative regret that this approach causes.

(b) After running 100 experiments with each parameter θ_i, A has obtained the following average performance results:

θ_1	θ_2	θ_3	θ_4	θ_5
0.5	0.8	0.9	0.6	0.5

Which parameter should A try out next? Explain your choice!
Hint: Use Hoeffding inequality with $T = 2$ and note that $\log(2) \approx 0.69$.