What is the value of future actions?

Readings for today

- Van Otterlo, M. (2009). Markov decision processes: Concepts and algorithms. Course on ‘Learning and Reasoning.

Topics

- Markov Process
- Markov Decision Process
- The Bellman solution

Markov Process

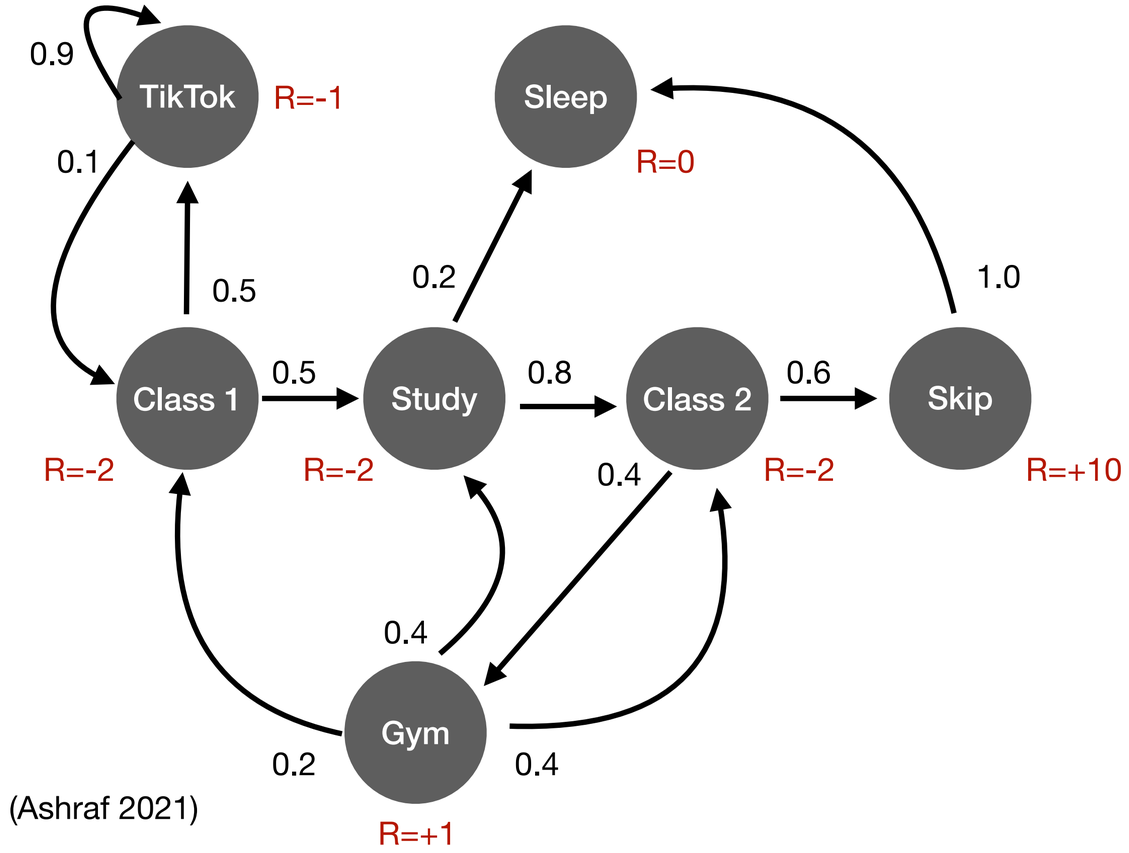
A day in your (college) life

What to do with your day?

- Go to class
- Skip class
- Study in the library
- Go to the gym
- Eat the UC
- Sleep in the dorm
- Play video games

Can only do one of these things at a time.
Each is associated with a particular “reward”

What is a Markov Process?



(Ashraf 2021)

Definition

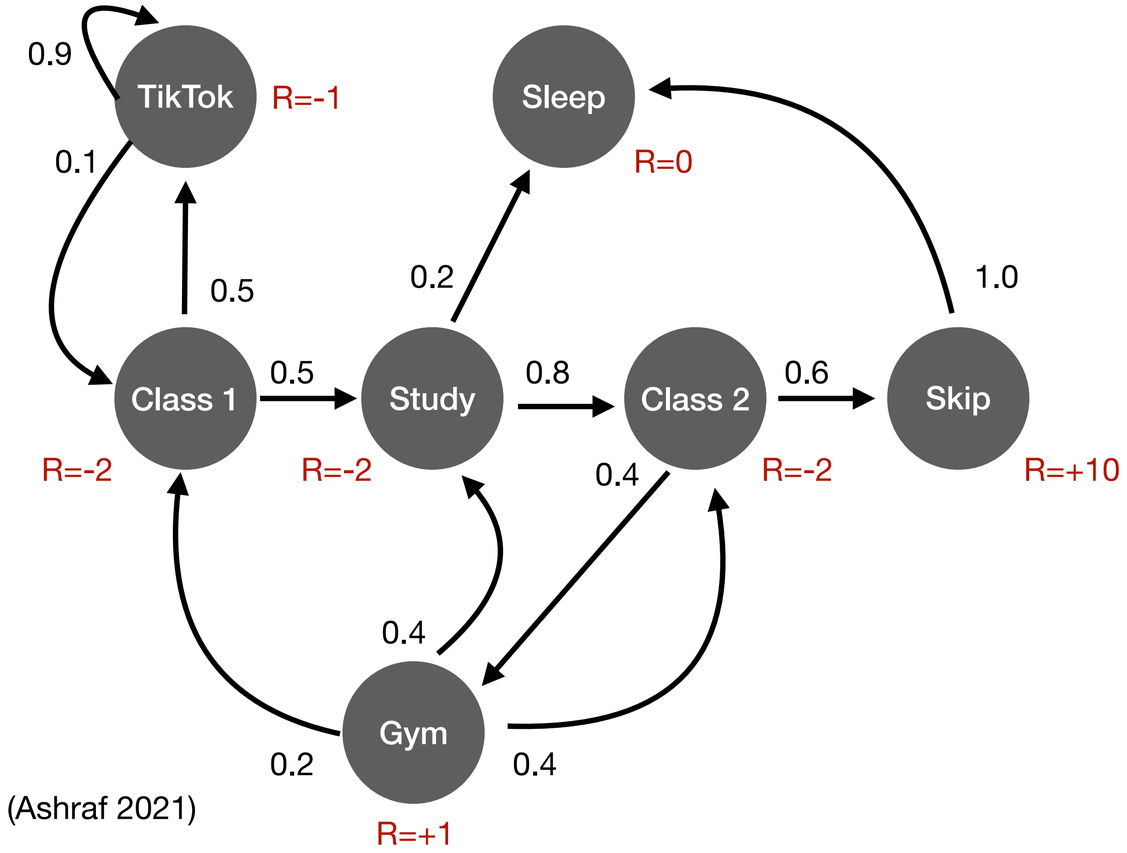
A stochastic (random) process that models a system moving between states based on transition probabilities.

Key components

- **States**: Represent all possible situations the system can be in.
- **Transitions**: Probability distributions that dictate state changes based on actions.
- **Rewards**: Numerical feedback given after state transitions.

(Van Otterlo 2009)

What is a Markov Process?



(Ashraf 2021)

All chance

Transitions between states happen according to fixed probabilities, without any external control or decision-making by an agent.

No memory

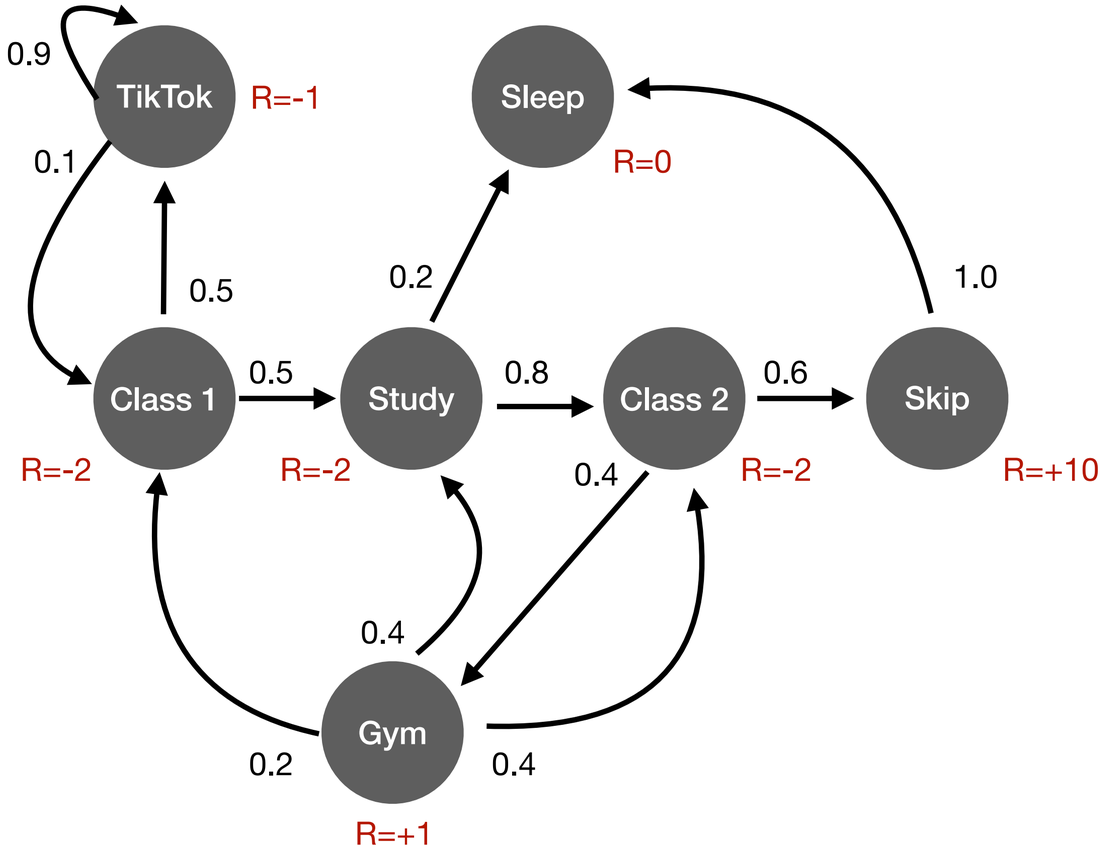
The future state depends only on the current state, not on the history of previous states.

↳ Markov property

(Van Otterlo 2009)

Markov Decision Process

What is a Markov Decision Process (MDP)?



Definition

Extends a Markov Process by adding decision-making capabilities. An agent can influence transitions by choosing actions in each state.

Key components

- **States**: Represent all possible situations the system can be in.
- **Actions (A):** Choices available to the agent in each state.
- **Transitions**: Probability distributions that dictate state changes based on actions.
- **Rewards**: Numerical feedback given after state transitions.

(Van Otterlo 2009)

Formal definition

Definition

an ordered, immutable collection of elements, a set

An MDP is formally defined as a *tuple*: $\langle S, A, T, R \rangle$

- **S:** A set of states $\{s_1, s_2, \ldots, s_N\}$
- **A:** A set of actions $\{a_1, a_2, \ldots, a_N\}$
- **T:** Transition function $T(s, a, s')$ reflecting the probability of moving to state s' after taking action a in state s.
- **R:** Reward function $R(s, a, s')$ reflecting the reward received after moving from state s to s' using action a.

How the environment changes in response to actions

How "good" or "bad" each action is, guiding the agent toward a goal.

Goal: Maximize the cumulative reward over time, driving the system toward optimal behavior.

(Van Otterlo 2009)

The Bellman solution

The Bellman equation

What is the *optimal* path through potential states that has the highest value?

$$v(s) = \mathbb{E}(G_t \mid S_t = s)$$

temporal discounting term

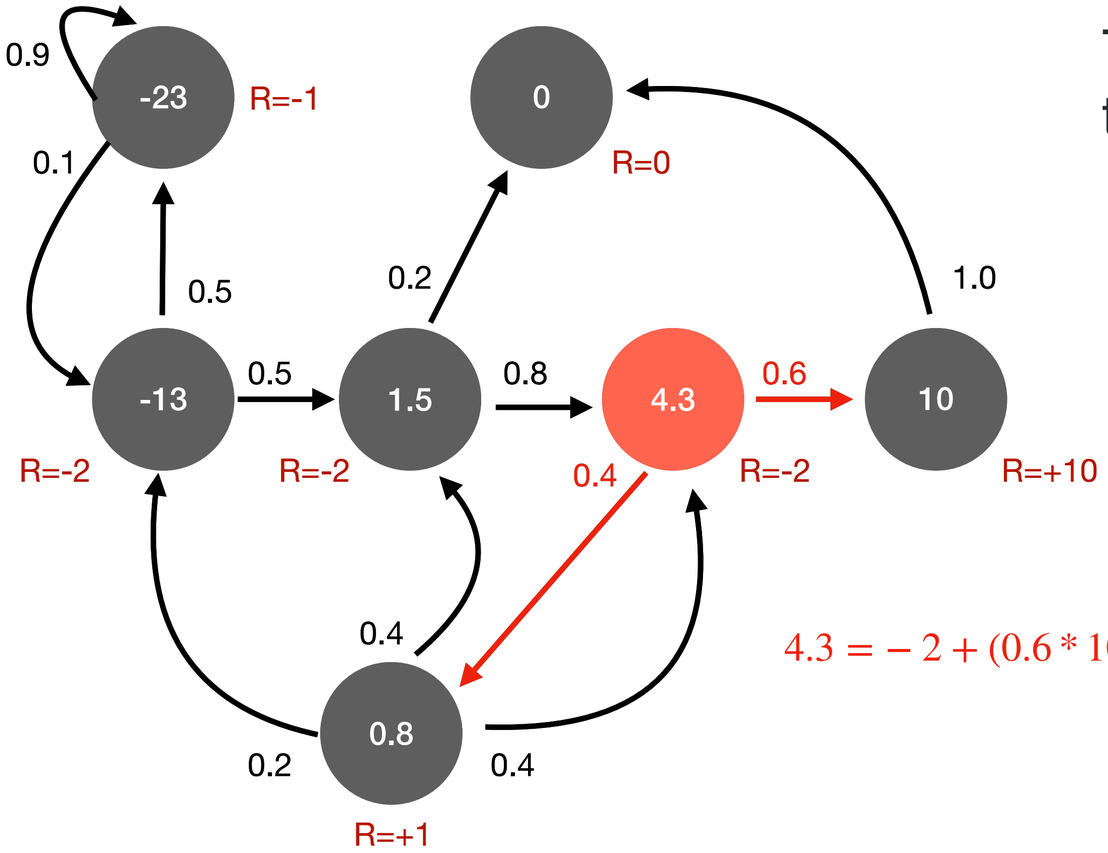
$$G_{t+1} \rightarrow v(S_{t+1})$$

Bellman equation

$$v(s) = \mathbb{E}(R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s)$$

(Ashraf 2021)

The Bellman equation

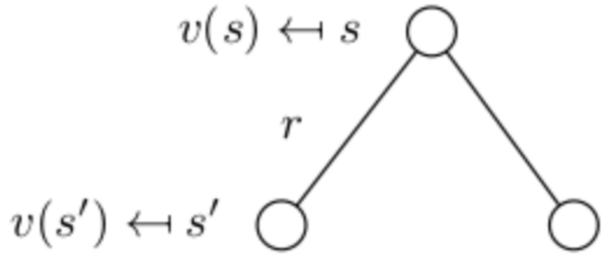


The value both depends on the reward and the transition probability .

$$v(s) = \mathbb{E}(R_{t+1} + \gamma v(S_{t+1}) \,|\, S_t = s)$$

$$= \mathbf{R}_S + \gamma \sum_{s' \in S} \mathbf{P}_{ss'} v(s')$$

$$4.3 = -2 + (0.6 * 10) + (0.4 * 0.8)$$



(Ashraf 2021)

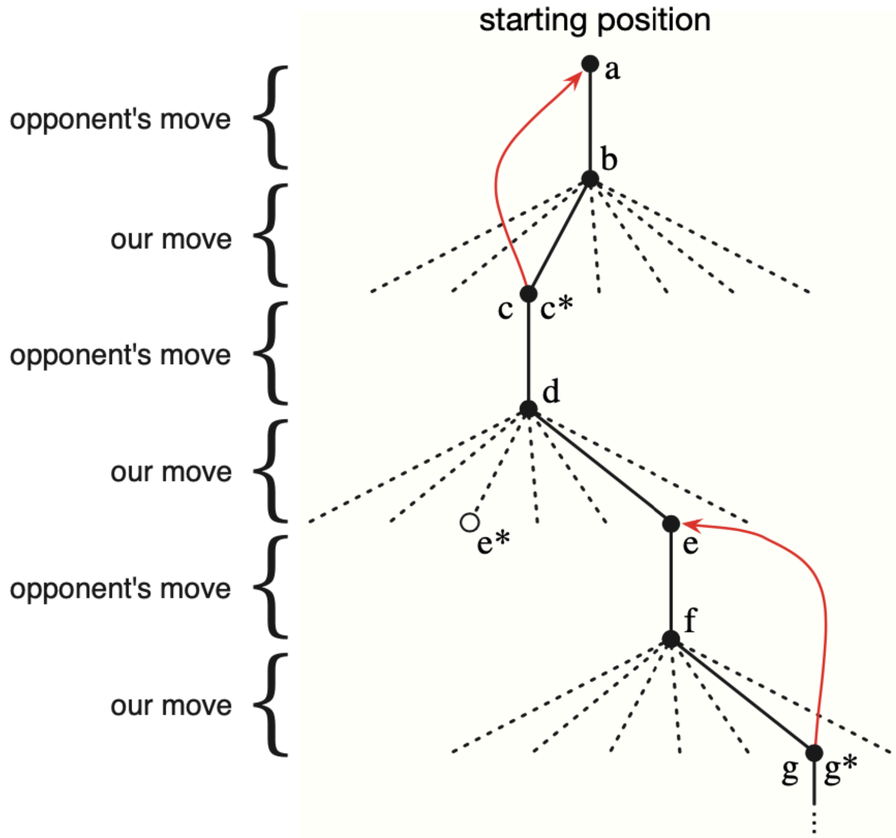
Temporal difference update

How do you update your state value from one move to the next?

$$V(S_{t+1}) \leftarrow V(S_t) + \alpha[V(S_{t+1}) - V(S_t)]$$

error

learning rate



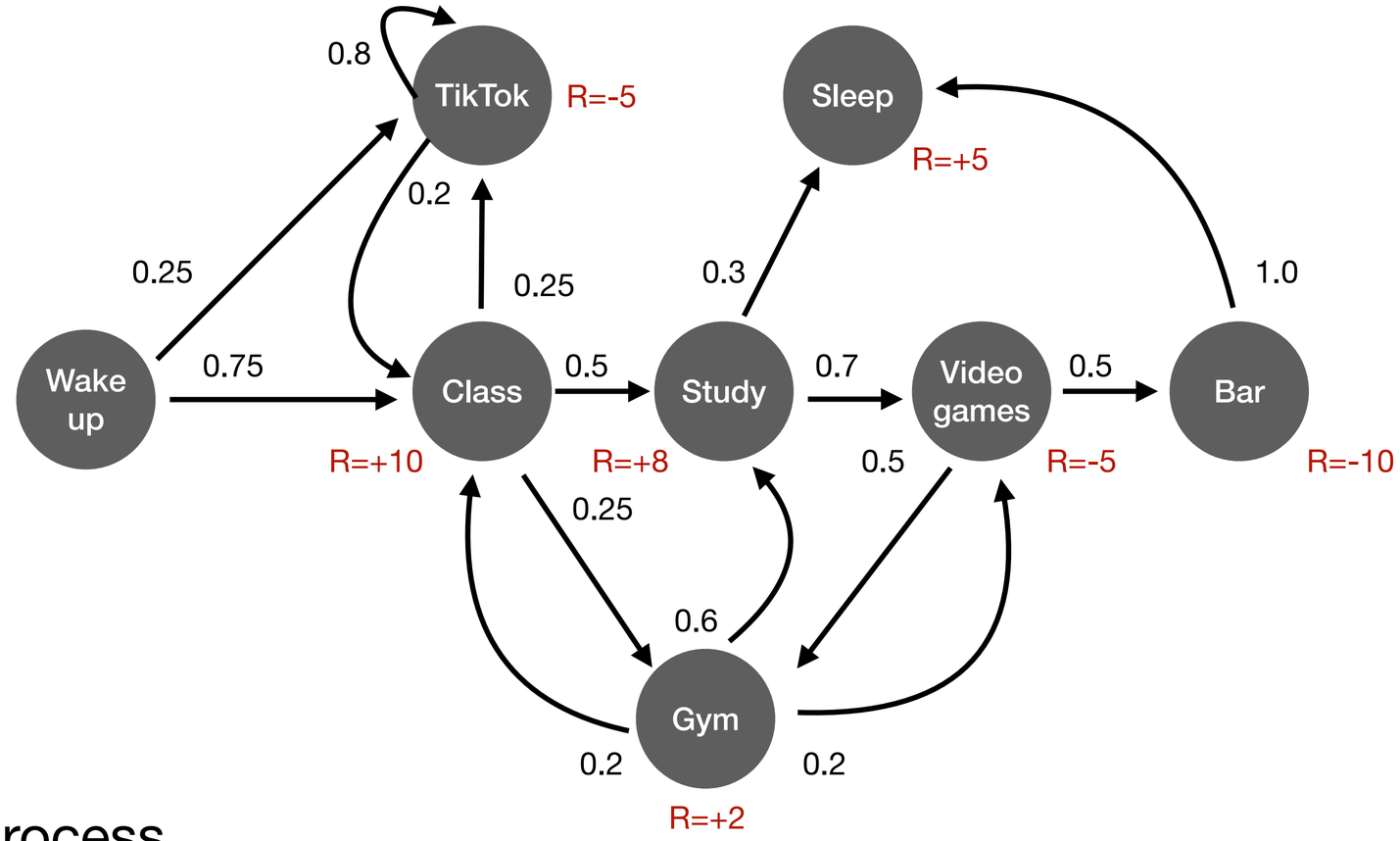
This works in the forward looking framing of the Bellman equation.

(Sutton & Barto, 2020)

Take home message

- Markov processes are a useful modeling framework for understanding sequences of memoryless decisions.
- Markov decision processes conceptually capture the structure of reinforcement learning.
- The Bellman equation is the locally optimal solution to identifying the next decision that maximizes future rewards.

The Markov Game



Goal:

Simulate your day as a Markov process

The Markov Game

Rules:

1. Start at the “Wake Up” state
2. Use the random number generator to determine which state to move to next following the Transition Guide.
3. Record the reward the next state you move into in the worksheet you were emailed.
4. Continue moving through the day until you get to the “Sleep” state ***or*** you complete 40 state transitions (maximum rows on the worksheet).
5. Record your total reward for the day.
6. Repeat steps 1-5 for 7 rounds.

Random Number Generator: https://www.random.org/

Min: 1
Max: 100

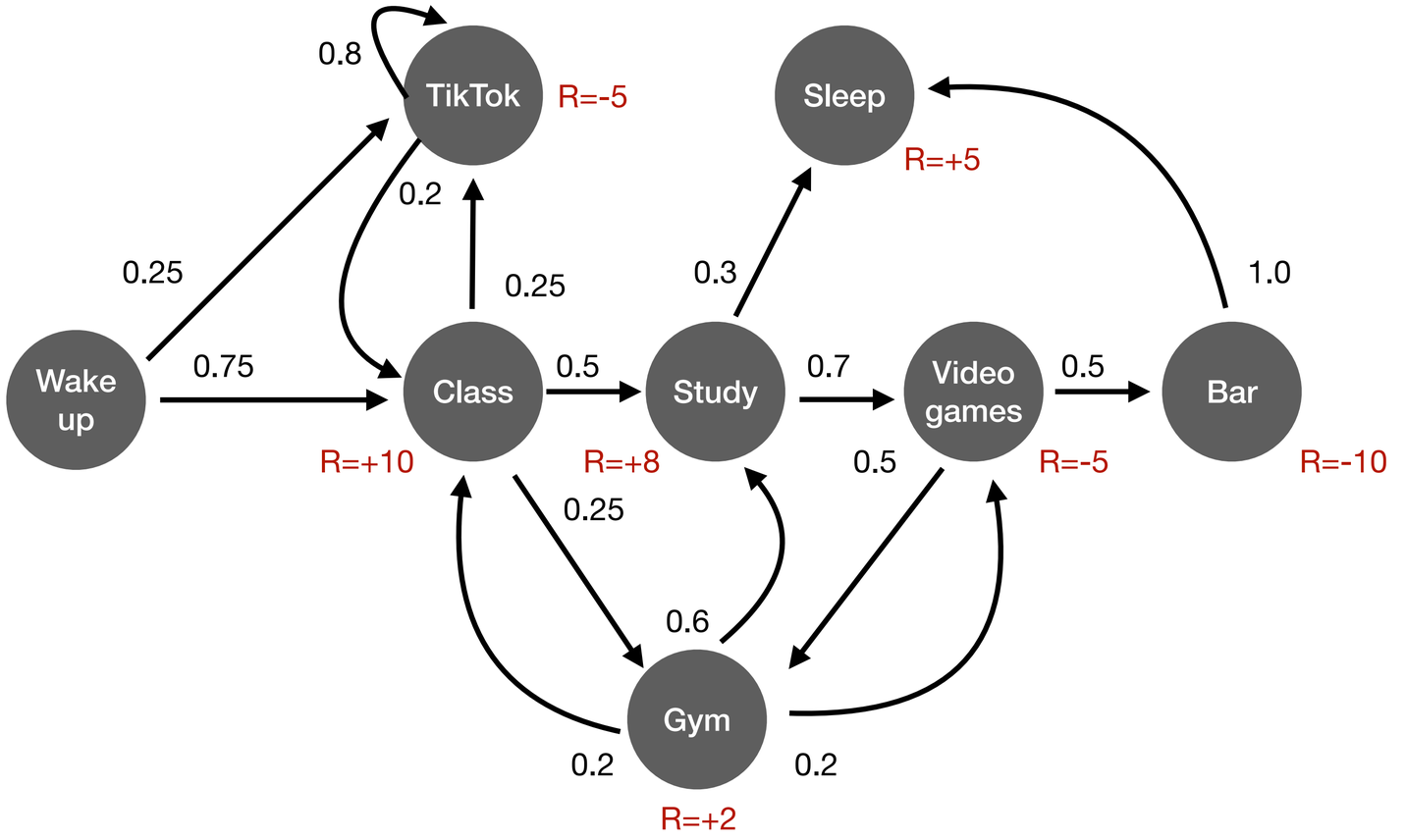
Transition Guide:

		Next State (s')						
		TikTok	Class	Study	Gym	Videogames	Bar	Sleep
Current State (s)	Wake up	<25	>=25					
	TikTok	>=20	<20					
	Class	<25		25-75	>75			
	Study					<=70		>70
	Gym		<20	20-80		>80		
	Videogames				<=50		>50	
	Bar							>0

Extra Credit (3 pts)

Task:

Use the Bellman equation to estimate the value of each state of the day. Submit answers by entering values in the last worksheet of the game spreadsheet file. Set $\gamma = 1$.



$$v(s) = \mathbf{R}_S + \gamma \sum_{s' \in S} \mathbf{P}_{ss'} v(s')$$

Due:

By 3pm (class time) Tuesday 10/31. Please submit by emailing to timothyv@andrew.cmu.edu