Week 8

Reinforcement Learning

- In what situations do we use reinforcement learning? What kinds of problems does it solve?
  - Give an example of RL used to play a game. Did it outperform humans? Does this scare you?
  - RL is useful for open-ended problems and more specifically those problems where it is easy to solve a problem using an agent-environment relationship like walking in a 3D world.
- What do the variables s, a, τ, represent in RL?
  - What do s', a' represent?
  - s is the state of the environment
  - a is the action taken by the agent
  - τ is the transition from s to s'
  - a' is the action taken by the agent in s'.
- What is an action space?
  - The possible actions that can be taken by the agent from a given state.
- What is an episode?
  - A series of actions taken by the agent in an environment.
- How is a neural network used to generate a probability distribution over actions given a state?
  - (i.e How would you construct a network if the dimensionality of the state was n1 and the action space had a cardinality of m1)
  - How is the log probability calculated from the logits of the NN output?
    - The log probability is calculated by taking the log of the softmax of the logits.
    - The input layer would have n1 neurons to read the state of the environment and the output layer will have a size of m1 as the probability distribution over actions, the most optimal action will be the one with the highest probability.
- Describe infinite-horizon discounted return and what the discount factor is
  - The infinite-horizon discounted return is the sum of all rewards obtained by the agent.
  - The discount factor is the rate at which the reward values diminish.
- Describe what a stochastic policy is (you may have to look up what stochasticity is), and what it means for a policy to be parameterized by θ
  - A stochastic policy is a policy where the agent has multiple actions to choose from. There is a probability distribution over actions given a state.
  - θ denotes the parameters of the policy which is optimized to find the best policy.

- What is a good policy looking to maximize?
  - Think: If you set up your rewards in your environment randomly, would an RL model learn anything? No. This ties into the greater idea of reward-shaping, which is the concept of how to place rewards in an environment to encourage "good" behavior by an RL agent.
  - Maximize the expected return.
- What is the difference between the value function and the Q-function?
  - Describe the connection between the two
  - The value function is the expected return given a state.
  - The Q-function is the expected return given a state and a policy to base actions on.
- Based on the idea behind Bellman Equations ("The value of your starting point is the reward you expect to get from being there, plus the value of wherever you land next.") Explain how the following two equations satisfy the idea behind Bellman Equations:
  - They both get the value from summing the reward and the discounted value of the next state.
- What is the advantage function?
  - The advantage function is the difference between the Q-function and the value function.

RL Algorithms

- What is the difference between model-free and model-based RL?
  - Model-free RL is when the agent does not have a model of the environment, their actions will be solely based on the policies and rewards.
  - Model-based RL agents can predict the reward for an action from a state, and then choose the most optimal action.
- What is the difference between policy approximators and Q-learning?
  - Policy approximators try to estimate the optimal policy by using a neural network to approximate the policy.
  - Q-learning is a model-free RL algorithm that tries to estimate the Q-function.
- Think back to what the $J(\pi)$ function was. How would performing gradient ascent (finding where this function is locally highest) on this function help the agent perform well in its environment?
  - $J(\pi)$ is the expected return given a policy, performing gradient ascent on this function will allow the agent to find the optimal policy to maximize the expected return.
- How is a policy derived (not literally a derivation) from the Q-function? Hint: argmax
  - The policy is derived by taking the action with the highest Q-value. Argmax is used to find the action with the highest Q-value.