# RBE 595 — Reinforcement Learning
# Assignment #6
# Model-Based Reinforcement Learning

**Arjan Gupta**

# Problem 1

What is "planning" in the context of Reinforcement Learning?

## Answer

In the context of Reinforcement Learning, planning is the process of using a model of the environment to improve the policy.

A model of the environment is a representation of the environment that is used to predict what next state and reward will be given a current state and action. Once we have a model, we can use it to simulate the environment and produce a simulated experience, in the form of an episode. Then we can use this simulated experience to improve the policy. This is the process of **planning**.

# Problem 2

What is the difference between Dyna-Q and Dyna-Q+ algorithms?

## Answer

The problem with Dyna-Q is that it does not balance exploration and exploitation well in the planning phase. This is because the planning phase is greedy. Only the learning phase is $\epsilon$-greedy. The reason that we want to also explore in the planning phase is that the model may need to change if the environment changes over time.

Dyna-Q+ is like Dyna-Q, except that it adds an exploration 'bonus' to the planning phase. What this means is that we essentially provide a bonus reward in the planning phase for states that have not been visited in a long time. This encourages the agent to explore in the planning phase. Specifically, the bonus reward is given by:

$$R = r + \kappa\sqrt{\tau(s,a)}$$

Where $r$ is the reward received, $\kappa$ is a constant of our choice, and $\tau(s,a)$ is the number of time steps since the last visit to state $s$ after taking action $a$. It is important to note that the bonus reward is only given in the planning phase, and not during regular interaction with the real environment.

# Problem 3

Model-based RL methods suffer more bias than model-free methods. Is this statement correct? Why or why not?

## Answer

This statement is correct. Model-based RL methods suffer more bias than model-free methods because the design of the model introduces bias. The model is a representation of the environment, and it is not possible to represent the environment perfectly. Therefore, the model will always introduce some bias.

# Problem 4

Model-based RL methods are more sample efficient. Is this statement correct? Why or why not?

## Answer

This statement is correct. Model-based RL methods can use a limited number of samples to learn a model of the environment. Given an episode from real-interaction, we can extract as much 'juice' as possible from it by using it to learn a model. Then, we can use the model to simulate more episodes, and use those episodes to improve the policy.

On the other hand model-free RL methods can only use the episode to improve the policy, so they require more 'samples' to learn the same amount.

Therefore, model-based RL methods are more sample efficient since they can learn more from the same number of samples.