

# RBE595 - Week 10 Assignment

Keith Chester

Due date: March 19, 2023

## Problem 1

*What is “planning” in the context of Reinforcement Learning?*

”Planning” in this context is the process of action selection in an environment while aiming to maximize a reward. Alternatively, ”learning” involves updating a value and/or policy function through observed experience.

## Problem 2

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

There are two notable differences. The first is that Dyna-Q+ includes an exploration bonus term in its update rule for its Q function. This results in an agent that is encouraged to explore states/actions that haven’t been explored yet, which is beneficial in non-static environments especially.

Another notable difference is that Dyna-Q+ utilizes a count-based mechanism to estimate the ”novelty” of a state/action pair; this is used to calculate the aforementioned exploration bonus.

## Problem 3

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

This statement isn’t correct. Model-free RL methods learn on-line actively interacting with the environment (which is necessary given the scenario, such as an environment that is too complex to simulate or is not understood at all). Alternatively model-based RL methods learn a model of an environment (or attempt to build a representation as best as possible) and from this tries to simulate the outcomes of different actions prior to actually interacting with the environment. If the model itself is a poor representation of the environment then it will introduce bias. If the model is a perfect or good-enough representation of the environment, however, this can be more efficient and less biased than an equivalent model-free approach.

## Problem 4

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

Yes, this statement is generally correct, depending on your environment and the accuracy of your representation of the environment. This is because you can quickly simulate additional training data/experience through the simulated environment and thus lowering the cost to sample the environment, as opposed to an agent that has to explore a real-world environment.

## Problem 5

*What are the 4 steps of the MCTS algorithm? How does MCTS balances exploration/exploitation?*

The four steps of the Monte Carlo Tree Search (MCTS) algorithm are selection, expansion, simulation, and backup.

**Selection** - Starting from the root node of the tree, a path is traversed through the tree to a leaf node using a selection policy. The selection policy balances exploration/exploitation at this stage by choosing either nodes with high value or high uncertainty (exploitation and exploration, respectively).

**Expansion** - On some iterations we create new children nodes added to the leaf node when the selection policy chooses unexplored options.

**Simulation** - Starting from the selected or resulting new child node we simulate a complete episode with the rollout policy.

**Backup** - The return generated by this episode is propagated ("backed up") through the tree updating its calculated value estimations.

## Problem 6

*The nonplanning method looks particularly poor in Figure 8.3 because it is a one-step method; a method using multi-step bootstrapping would do better. Do you think one of the multi-step bootstrapping methods from Chapter 7 could do as well as the Dyna method? Explain why or why not*

A multi-step bootstrapping method (MSBM) likely will not do as well as the Dyna method. MSBM have to experience the environment to learn its value estimation function, but it's model-free and thus can't leveraged the learned model of the environment in order to bootstrap with additional training data. The Dyna method, by contrast, has higher sample-efficiency and is generally more effective in complex and/or stochastic and/or changing environments.

## Problem 7

*Why did the Dyna agent with exploration bonus, Dyna-Q+, perform better in the first phase as well as in the second phase of the blocking and shortcut experiments?*

The exploration bonus encourages the agent to visit states that were previously unvisited. This increased emphasis on exploration means the agent has a better idea of the environment and can better find optimal solutions to a given environment. Dyna in contrast doesn't explore as much and can become stuck on a set solution in examples where the environment is stochastic.