# Written Assignment 4

Due: Friday 12/05/2025 @ 11:59pm EST

## Disclaimer

I encourage you to work together, I am a firm believer that we are at our best (and learn better) when we communicate with our peers. Perspective is incredibly important when it comes to solving problems, and sometimes it takes talking to other humans (or rubber ducks in the case of programmers) to gain a perspective we normally would not be able to achieve on our own. The only thing I ask is that you report who you work with: this is **not** to punish anyone, but instead will help me figure out what topics I need to spend extra time on/who to help. When you turn in your solution (please use some form of typesetting: do **NOT** turn in handwritten solutions), please note who you worked with.

**Question 1: The Precariousness of RL (25 points)**

In the last written assignment you considered a loss function where each point was weighted with a nonzero coefficient $r^{(i)}$. You showed that the optimal solution was a function of the weights each sample was given. In this problem you will express an arbitrary supervised learning dataset (which we use in RL) into a dataset where samples are given weights:

1. Whenever we have a dataset where each sample is unique (e.g. does not appear more than once), we are creating a dataset where each sample is given a uniform weight. Show that when a dataset contains duplicate points we are creating a dataset where samples are not given uniform weights.

2. Now relax the weighting terms so that every weight $r^{(i)} \geq 0$ instead of $> 0$. Show that this loss function over a dataset of finite samples can be converted into a weighted sum over all possible data points.

3. What happens to terms with weights of 0? Do they impact the solution at all? How do we know that the model will perform well on these points?

**Note:** this is a **proof** question, meaning you must follow formal proof structure (see the examples of piazza for guidance). There is one exception: you do not have to follow formal proof structure when answering the last part, natural language is fine but I will not accept hand-wavy justification.

**Question 2: Reward Function Flavors (25 points)**

In lecture, we talked about MDPs that are formulated with a reward function $R(s)$ (i.e. the reward only depends on the current state). However, sometimes MDPs are formulated with a reward function $R(s, a)$ (i.e. a reward function that depends on the action taken), or even $R(s, a, s')$ (i.e. a reward function that depends on the action taken and the way the action is resolved). In this problem, you will show that even though someone may choose one flavor of reward function over another, they are actually identical:

1. Write the bellman equation that uses $R(s, a)$ and write the bellman equation that uses $R(s, a, s')$
2. Show how an MDP with reward function $R(s, a, s')$ can be converted into a different MDP with reward $R(s, a)$ such that optimal policies in the new MDP correspond exactly to optimal policies in the original MDP.
3. Show how an MDP with reward function $R(s, a)$ can be convered into a different MDP with reward $R(s)$ such that optimal policies in the new MDP correspond exactly to optimal policies in the original MDP.

**Note:** this is a **proof** question, meaning you must follow formal proof structure (see the examples of piazza for guidance).

### Question 3: Sum of Discounted Rewards vs. Max Reward (25 points)

In lecture we defined the utility of a trajectory to be some additive combination of the rewards along that trajectory. So far this has taken two forms: additive rewards and discounted rewards. However, what happens if we define the utility of a trajectory as the maximum reward observed in that trajectory? Show that this utility function does not result in stationary preferences between trajectories (i.e. that such an agent may change its preference for the optimal trajectory as a function of time). Is it still possible to define autility function on trajectories such that a policy which maximizes the expected trajectory utility results in optimal behavior?

**Note:** this is a **proof** question, meaning you must follow formal proof structure (see the examples of piazza for guidance).

## Question 4: Proof that the Bellman Equation is a Contraction Function (25 points)

In lecture we claimed that the bellman equation is a contraction function. Specifically, we said that, for any two vectors of utilities $\vec{u}$ and $\vec{u}'$:

$$||B(\vec{u}) - B(\vec{u}')||_\infty \leq \gamma||\vec{u} - \vec{u}'||_\infty$$

1. Show that, for any functions $f$ and $g$:

$$|\max_a f(a) - \max_a g(a)| \leq \max_a |f(a) - g(a)|$$

2. Derive an expression for $\left|\Big(B(\vec{u}) - B(\vec{u}')\Big)(s)\right|$ and then apply the result from part 1 to complete the proof that the bellman equation is a contraction function.

**Note:** this is a **proof** question, meaning you must follow formal proof structure (see the examples of piazza for guidance).