Question 1. (8 points)

Consider a Markov decision process.

1. (3 points) Explain why a fixed (independent of $x_1, x_2, \ldots$) sequence of actions as $y_1, y_2, \ldots$ ($y_k \in Y$) does not solve a Markov decision process, i.e. cannot guarantee optimality in reinforcement learning. In which field of AI would a fixed sequence of actions be an appropriate solution? Where is the boundary between game theory and reinforcement learning?
2. (2 points) Recall the value iteration algorithm. This algorithm is based on a general method for solving a set of equations. Give the name for this method and explain how it works in one or two sequences.
3. (3 points) Recall the policy iteration algorithm. This algorithm is based on another well known concept from machine learning and statistics. Name this algorithm and explain its idea in one or two sequences. Provide an example of other usages of this algorithm in computer science or mathematics.

Question 2. (5 points)

We state the Bellman equations the following way:

$$U(x) = r(x) + \gamma \max_{y \in Y(x)} \sum_{x'} P(x' \mid x, y) U(x').$$

In some literature, you may find under the same name a different equation:

$$U(x) = \max_{y \in Y(x)} \sum_{x'} P(x' \mid x, y)(r(x, y, x') + \gamma U(x'))$$

Describe in natural language the difference between the two formulations and decide if they are equally general, or else describe which one is the more general and why.

Question 3. (12 points)

Despite many reinforcement learning algorithms with additive rewards, it is common to use discounted rewards to model the environment.

1. (1 point) Give the range for the discount factor parameter.
2. (1 point) How does the agent behave when $\gamma = 0$?
3. (2 points) Give an example of an environment where a high discount factor is a good choice and why. Do the same for a low discount factor.
4. (3 points) In the case of the infinite horizon, discounting the rewards is necessary. Explain why.
5. (3 points) Imagine that your fancy reinforcement learning algorithm is not working on a chess game. Is it a good idea to include the discount factor in grid-search on meta-parameters of your algorithm? If yes, explain why; if not, would you still consider a range of discount parameter values and why?
6. (2 points) Suppose that
$$\max_{x \in X} |r(x)| = r_{\mathsf{max}}.$$
Using only r_{max} and γ, give the tightest lower and upper bound on the cumulative discounted reward in a single episode.