# Notes on Policy Classes in Reinforcement Learning

based on the discussion in Wouter van Heeswijk’s [Four Policies of Reinforcement Learning](https://towardsdatascience.com/the-four-policy-classes-of-reinforcement-learning-38185daa6c8a),

written by D. Gueorguiev, 12/15/23

Assumption: We formulate the Reinforcement Learning problem as Markov Decision Process (MDP) model.

**Markov Decision Process** (abbrev *MDP*): a 4-tuple with

* is a set of states (finite or infinite)
* is a set of actions (finite or infinite)
* is the probability to get from state to state with action and with reward .
* is the discount factor which determines to what extent the focus is on the most recent rewards. with there is no focus on the most recent rewards only.

**Learning Policy** (or just *Policy*): function which represents mapping from states to probabilities of selecting each possible action.

If the agent is following policy at time , then is the probability that if . Note that is an ordinary function which defines a probability distribution over for each .

**Bellman equations**:

for all (1)

(2)

(3)

(4)

(5)

(6)

The goal is to solve the corresponding system of Bellman equations and thereby find the optimal policy .

Wouter van Heeswijk here references the following variant of the Bellman equation:

(7)

Here with