# Reinforcement Learning
# Exercise 03

Finn Hülsmann & Dennis Heinz

**Task 1.**

We created a Feed-Forward Neural Network with observation space sized input, one hidden layer with variable size and ReLU activation functions, and an output layer, with as many neurons as there are actions in the respective environment.

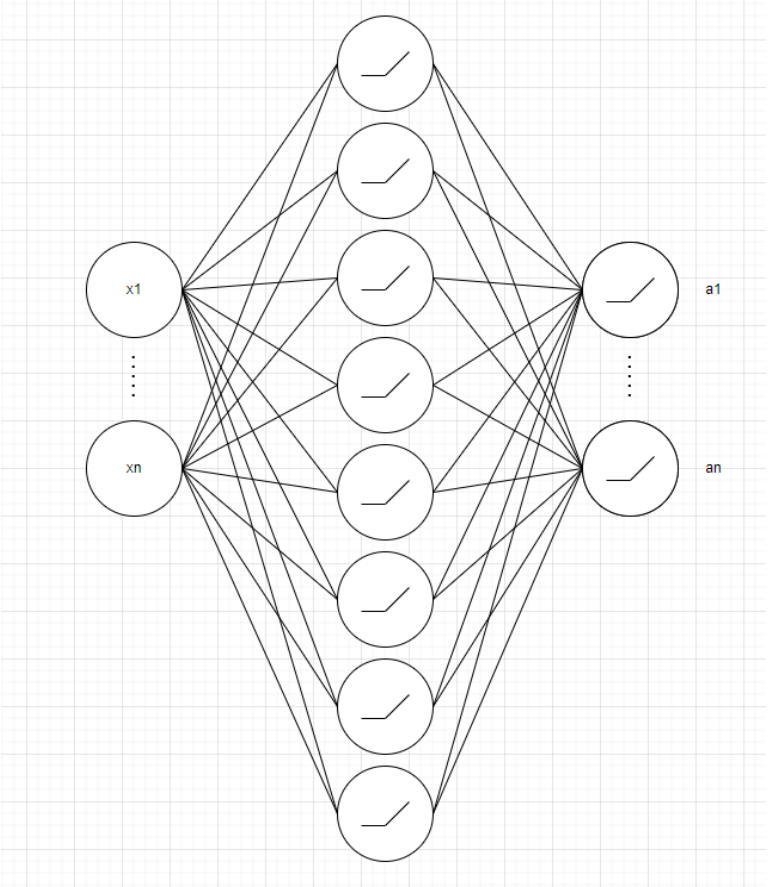


**Figure 1: ANN Architecture**

**Task 2.**

In the tabular case, a specific prediction objective was not needed. The learned value function matches the true value. In the non-tabular case, due to the generalization impact, we need to define an accuracy metric on the entire state space. For that, we use MSE for action values as the objective function. The goal is to find a $\boldsymbol{w}$ * that minimizes the MSE. Since we do not know the true target, we have to estimate it by using the sample return

accordingly to Monte Carlo (MC). Therefore, the true target $q_\pi(s_k, a_k)$ is approximated by the full episodic return g as $q_\pi(s_k, a_k) \approx g$. We focus on the on-policy case.

We noted that for different seeds, the resulting models can differ vastly in their performance in navigating in the environment. Parameters:

cartpole:
- epsilon = 0.1
- nr_episodes = 20000
- max_t = 400
- gamma = 0.9999
- replay_buffer_size = 10000

Mean episode reward: 81.2

mountaincar:
- epsilon = 0.1
- nr_episodes = 20000
- max_t = 4000
- gamma = 0.9999
- replay_buffer_size = 10000

Mean episode reward: -4000.0

**Task 3.**

For this task, we have filled in the respective code snippets. During the subsequent training, it was determined that the performance strongly depends on the selected hyperparameters. Also good models became worse over time again. Something like early stopping might help. For the “cartpole”, no suitable hyperparameters could be found, despite the fact that numerous combinations were tested. For the “mountaincar”, on the other hand, good results were achieved, although these depended strongly on the seed. Thus, very good models can be created, but also very bad ones, using the same hyperparameters.

The following hyperparameters were used:

cartpole:
- gamma=0.9999,
- epsilon=0.2,
- nr_episodes=1500,
- max_t=4000,
- warm_start_steps=4000,
- sync_rate=256,
- replay_buffer_size=5000,
- train_frequency=8,
- batch_size=128

Mean episode reward: 9.2

mountaincar:
- gamma=0.9999,
- epsilon=0.05,
- nr_episodes=1500,
- max_t=4000,
- warm_start_steps=4000,
- sync_rate=256,
- replay_buffer_size=5000,
- train_frequency=8,
- batch_size=128

Mean episode reward: -149.6