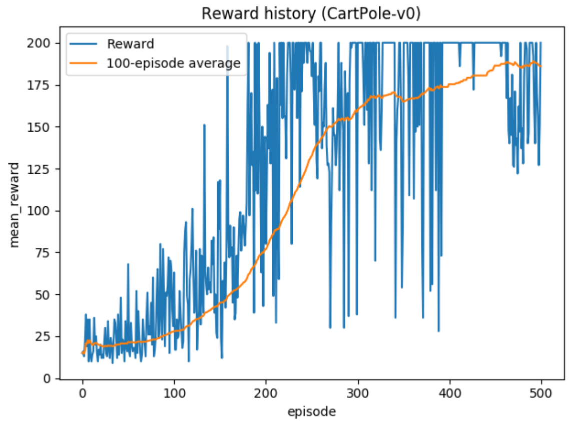
Reinforcement Learning

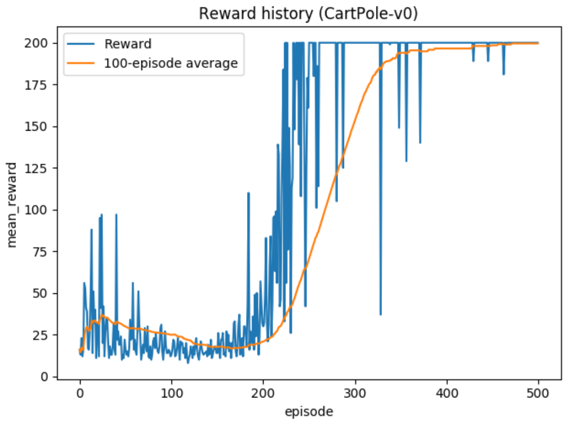
Exercise 1

Q1: Can the same model, trained with 200 timesteps, balance the pole for 500 timesteps? Why/why not?

It depends on the training process. If the agent was able to learn such a good policy, it could be that it manages to balance the pole for 500 timesteps. But it could also happen that the learned policy is not enough to do so, the average test reward could even be lower than 200.

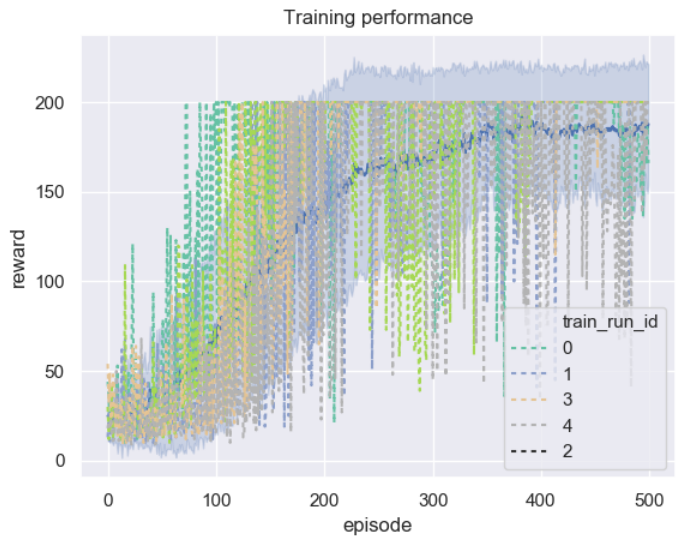
The following images show an example of bad and good training in terms of reward history during training. The first model got just 174 mean reward during test, while the latter got 500.





Q2: Why is this a case? What are the implications of this, when it comes to comparing reinforcement learning algorithms to each other?

The trade-off made between exploration and exploitation can create a great variance between successive runs of the same training process.



This implies that comparisons should be performed carefully: for each algorithm to be compared, either you use a long training process in order to converge to optimal policy, or you consider an average performance on multiple experiments.

For the considered algorithm, an overview of 100 independent executions can be seen in the following image.

Q3: How does changing the reward function impact the time needed for training?

A more complex reward function may require additional episodes, i.e. more time, to be learned, since the algorithm should explore more to fit the function in a proper way. That is why I ran my executions on the three new reward functions with 1000 episodes, instead of the default value 500.

Additional material

Under the folder */Images* there are graphs regarding reward history and multiple independent executions training for each of the reward functions. The corresponding models, again divided by reward function, can be found in the folder */Models*.