n-step Bootstrapping

Exercise 1

- *The one-step method strengthens only the last action of the sequence of actions that led to the high reward, whereas the n-step method strengthens the last n actions of the sequence* (Sutton and Barto) Similarly, a simulated sequence of 50 one-step updates from the Dyna-Q model also strengthen 50 actions. So for a 50-step bootstrapping we can expect similar behaviour. The difference is that in one direct RL step, Dyna-Q can sample different paths and not a single one. So probably n-step Off-policy Learning by Importance Sampling is what we need.
- Because DynaQ+ has better exploration strategy. It doesn't explore on costs of doing sub-optimal actions, but is still able to simulate better episodes if the dynamics have changed and a better model exists.

Exercise 2

The true values were computed as an average of 30 runs of TD(0) from last exercise.