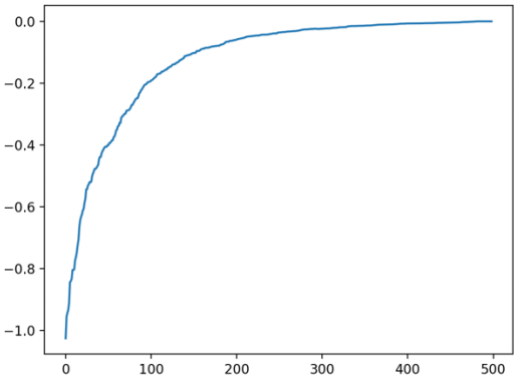
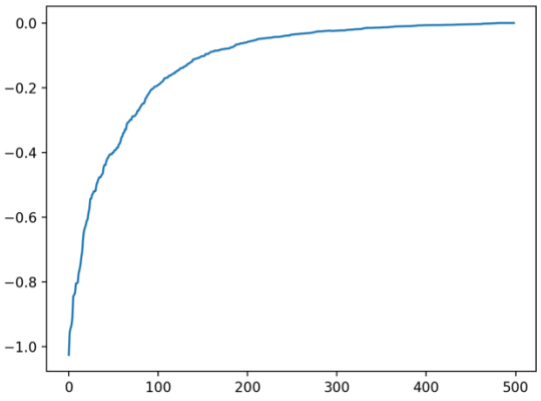
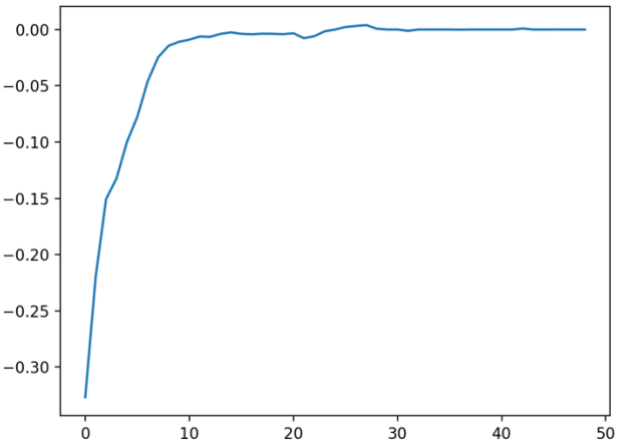
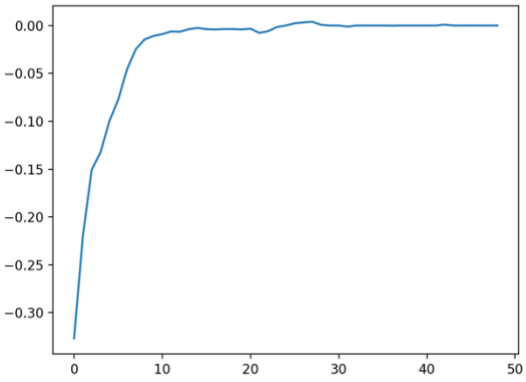
- Dyna vs non-Dyna
    - How does the final median reward compare? What about intermediate rewards at various trip numbers?
    - Does Dyna improve the speed of learning accurate Q values in terms of number of trips/experiences?
    - Do you see any downsides or problems with Dyna?





The above plot on the left shows the results for a non-Dyna learner with 500 trips where the x-axis is the trip number within a trial and the y axis is the median difference between each state Q value and the final Q value. The right plot takes the median across all trials to eliminate noise.



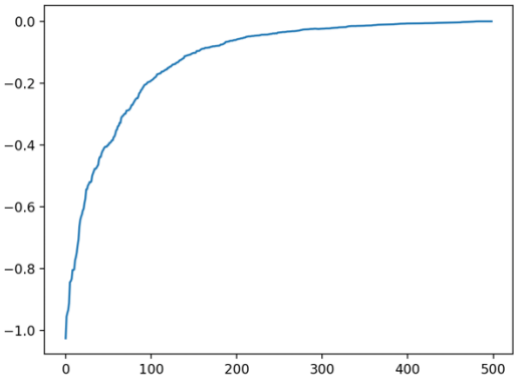


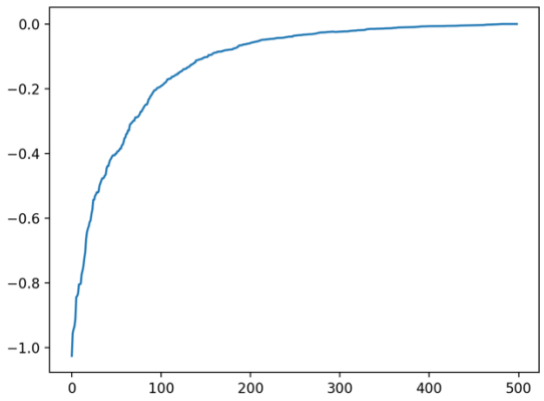
The above plot on the left shows the same information as the previous plots but instead for a Dyna-Q learner with 50 trips and 200 hallucinations per real experience.

The final median reward for the non-Dyna learner was -34.0, compared to -32.0 for the Dyna learner. As for intermediate rewards, the Dyna learner reaches close to its best reward in far fewer trips than the non-dyna learner, roughly 10 trips compared to 200. This suggests Dyna improves the speed of learning accurate Q values in terms of number of trips/experiences. This makes sense given for each real experience the dyna learner is simulating 200 additional experiences to learn from, so it is unsurprising it reaches more accurate Q values per experience.

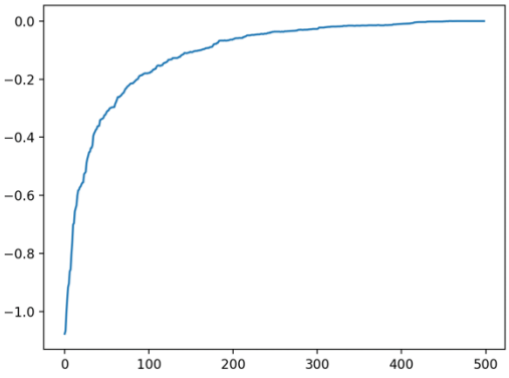
As for downsides of Dyna, there are increased memory demands because my implementation maintains a list of all previous experiences from which I can sample from when hallucinating.

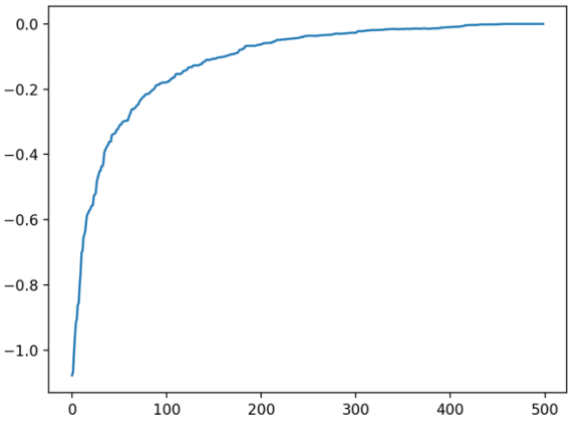
- Double Q vs Tabular Q
    - Does Double Q seem to improve the learning process? Does it converge in fewer trips?
    - Does Double Q seem to reduce overestimation bias for our problem?
        - If so, support with evidence. If not, give a logical explanation why our problem is different from the Double Q paper.





The above plots show the results for a Tabular Q learner with 500 trips (same axis as the above plots).



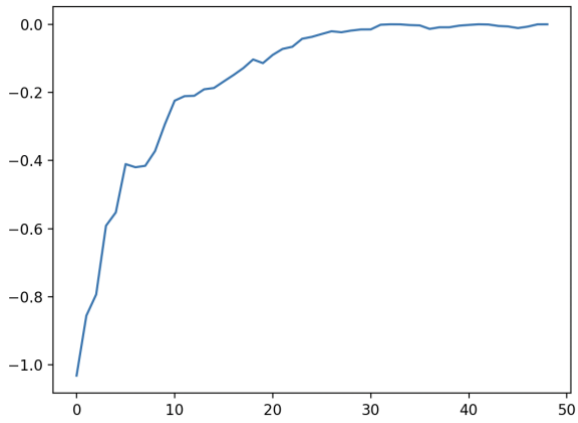


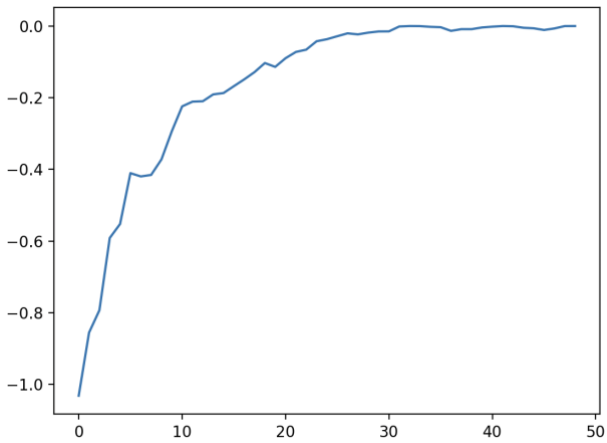
The above plots show the results for a Double Q learner with 500 trips.

Double Q doesn't seem to have a huge impact on the learning process compared to the Tabular Q learner. However, it does seem to move towards the final reward value quicker as we can see after 100 trips of the Double Q learner we are at about -0.18 compared to -0.2 for the Tabular Learner. This suggests the Double Q learner does converge in fewer trips.

The double Q learner does not seem to reduce overestimation bias since both curves are of very similar shapes. This is because the two Q-value estimators in Double Q-learning may become correlated during training causing the double Q learner to behave like a Tabular Q learner.

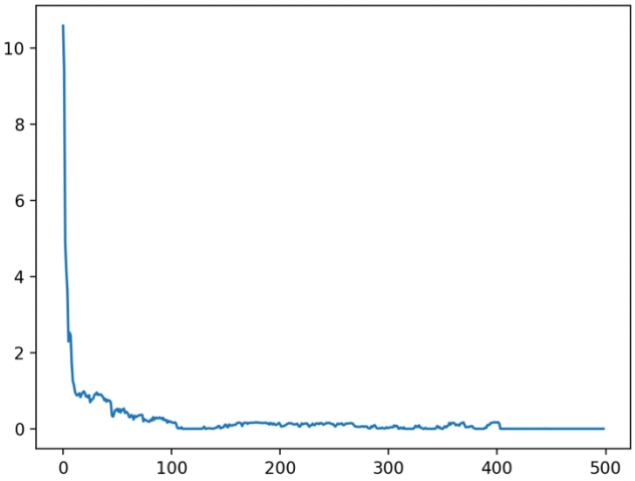
- Does there seem to be any benefit in combining Dyna and Double Q?



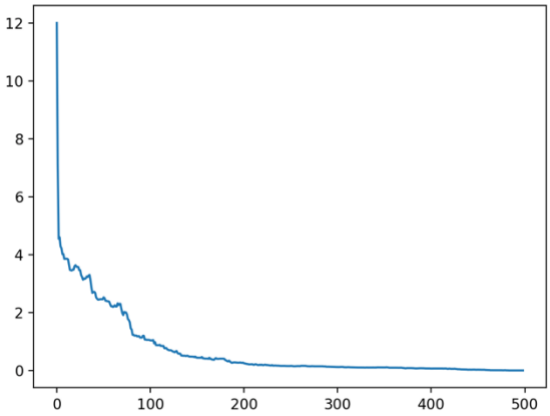


The above plots show the combination of Dyna and Double Q where the dyna parameter is 200. Compared to the regular Dyna learner, the combination of learners does not seem to impact the speed of learning accurate Q values in terms of number of trips/experiences. There seems to be less overestimation thanks to the Double Q learner component as indicated by the more gradual climb as opposed to steep increase and then flattening of the curve seen for the Dyna learner.

- Theoretically, the TIME_PENALTY of -1 shouldn't be doing anything, since we have a gamma value discounting future rewards.
    - Try removing the penalty and see what happens. Does learning work better, worse, about the same?
    - Does this change the functioning of Double Q with respect to overestimation?



This is a plot for a tabular Q learner with the time penalty set to 0. This had a huge impact, improving the final mean reward to 1.0 from -30.0 and saw a quicker convergence to this final value after about 100 experiences instead of 200.



As for the impact of removing the time penalty on a double Q learner we can see that the intermediate Q values initially greatly overestimate the median reward but that overestimation decreases with experiences.