

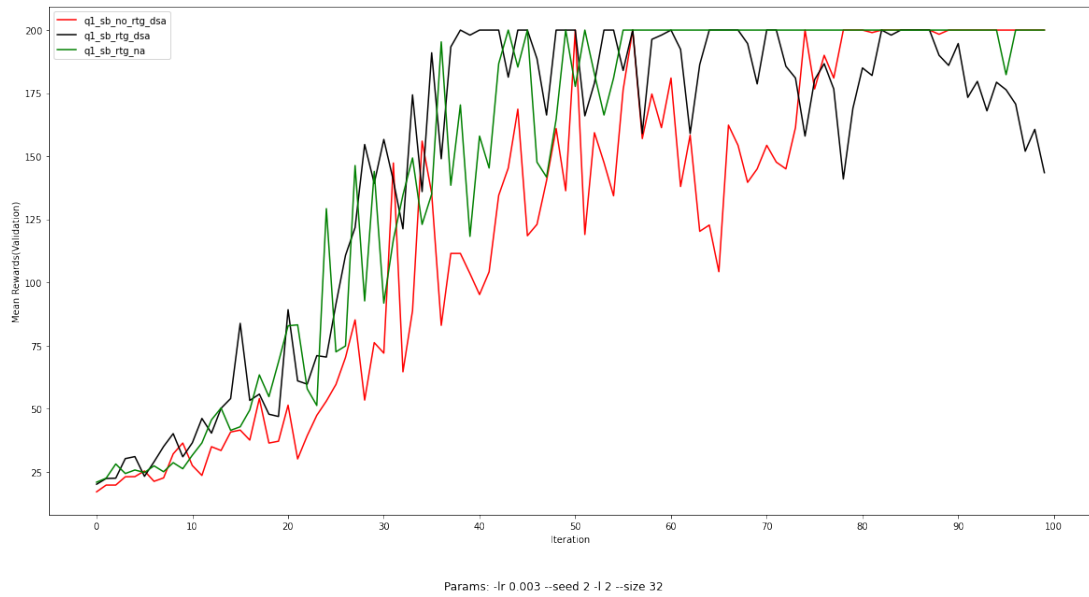
# Assignment 2: Policy Gradients

huseyinabanox@gmail.com

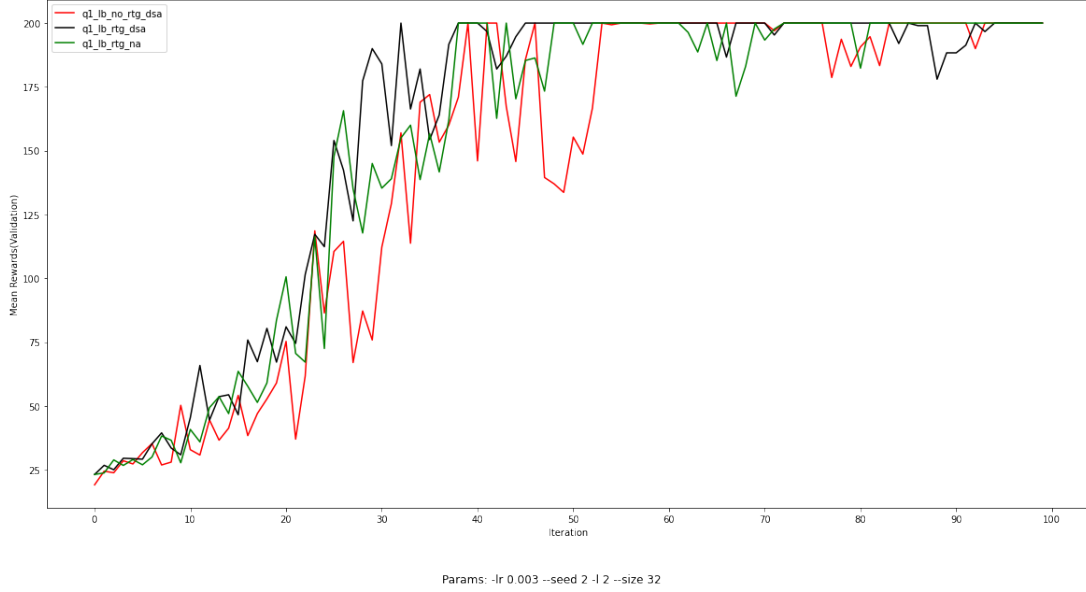
January 2023

## 1 Small-Scale Experiments

### 1.1 Experiment 1

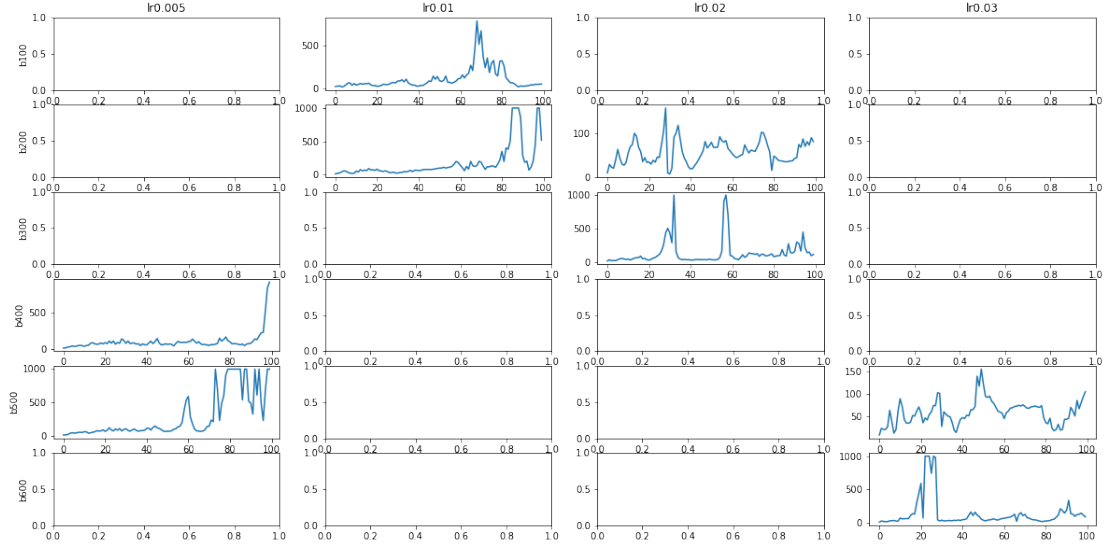


Reward-to-go estimator performs better. Advantage standardization difference is not statistically significant.



Reward-to-go estimator performs better. Advantage standardization difference is not statistically significant. Increasing the batch size helps the trajectory-centric algorithm converge with fewer iterations.

## 1.2 Experiment 2

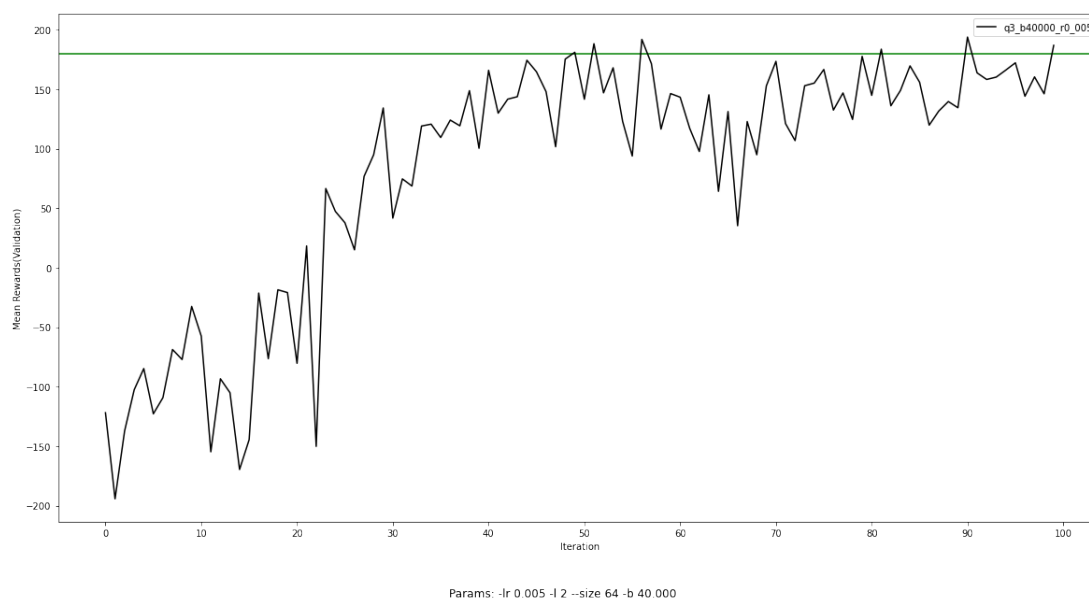


A grid search with default parameters shows that it is possible to reach the target reward of 1000 with learning rate 0.01 and batch size 200 or with

learning rate 0.02 and batch size 300. The first options seems to be more stable.

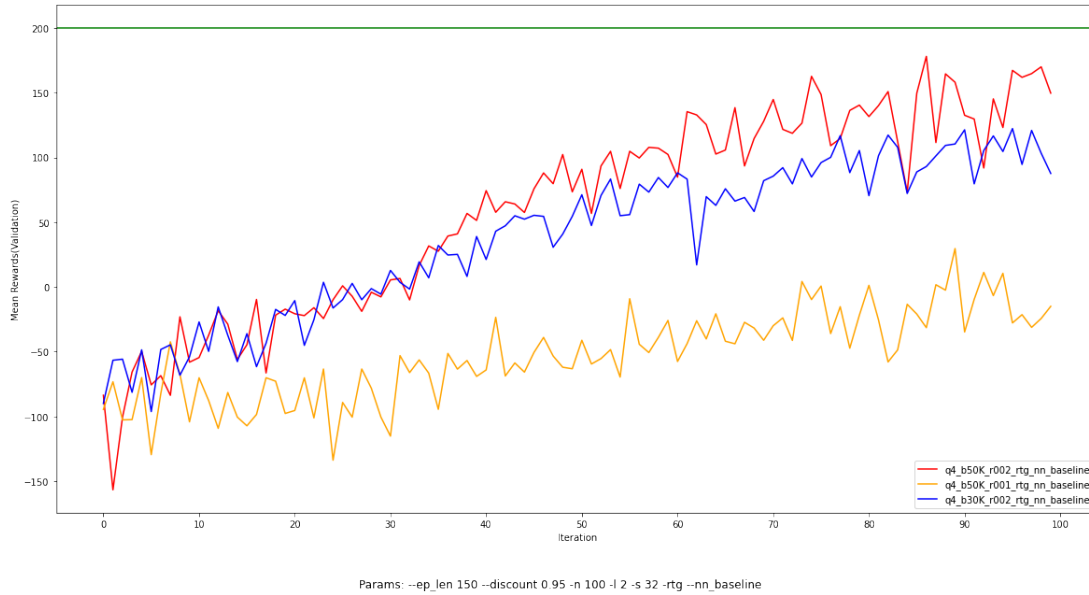
Please note that the figure tries to show the boundary in hyperparameter space. In each column the first chart indicates the largest learning rate and smallest batch size that fails to reach 1000. In each column the second chart indicates the largest learning rate and smallest batch size that succeeds in reaching 1000.

### 1.3 Experiment 3



### 1.4 Experiment 4

A trial and error search is run on hyperparameter space. Learning rate 0.02 and batch size 50K are found to be best performing.

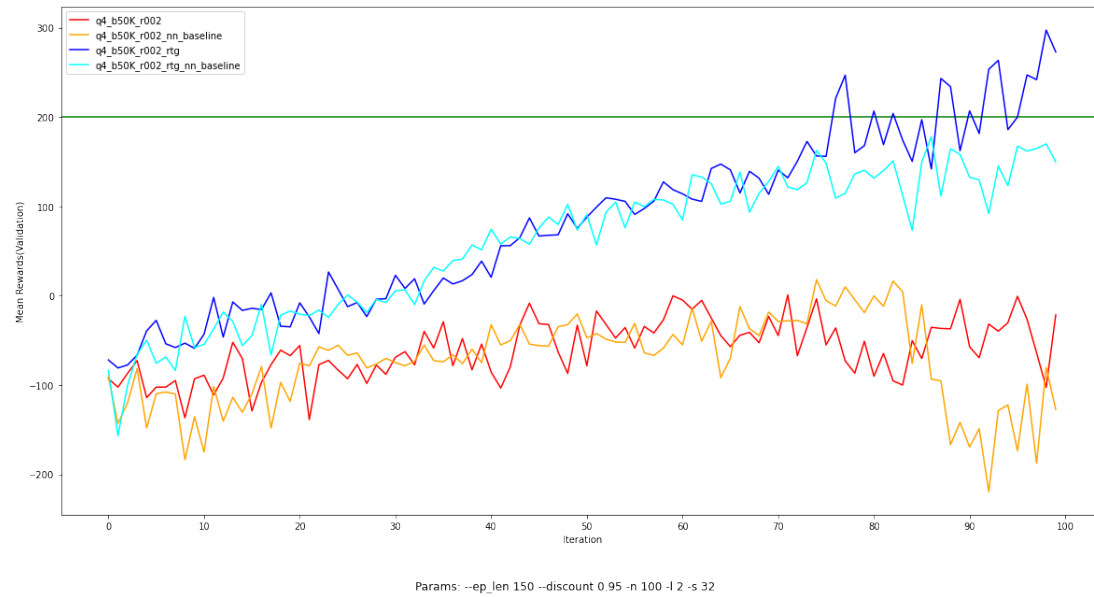


The chart above shows evaluation results by iteration for different parameter sets.

Bigger batch size is better. It reduces variance and speeds up learning.

Bigger learning rate is better. It speeds up learning. If learning rate is too large, it may cause converging to a local minimum or divergence.

While doing trial and error, largest possible values are selected at first. Then a smaller learning rate configuration is tested. Then a smaller batch size configuration is tested. Both configurations are found to perform worse than the first test.



The chart above shows performance for different reward-to-go and base-line parameter configurations. Reward-to-go always performs better.