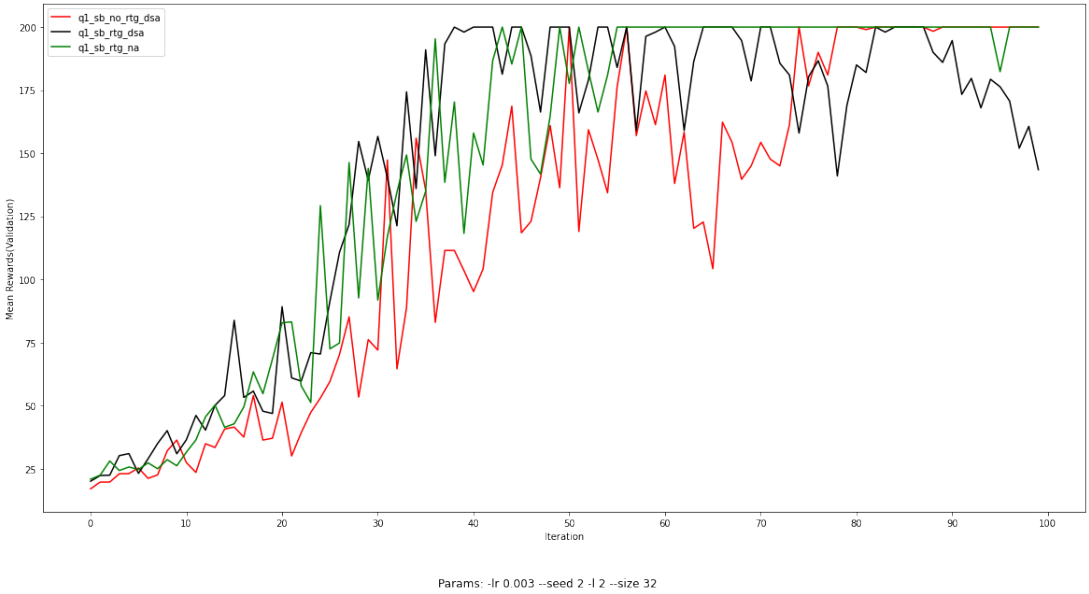
Assignment 2: Policy Gradients

huseyinabanox@gmail.com

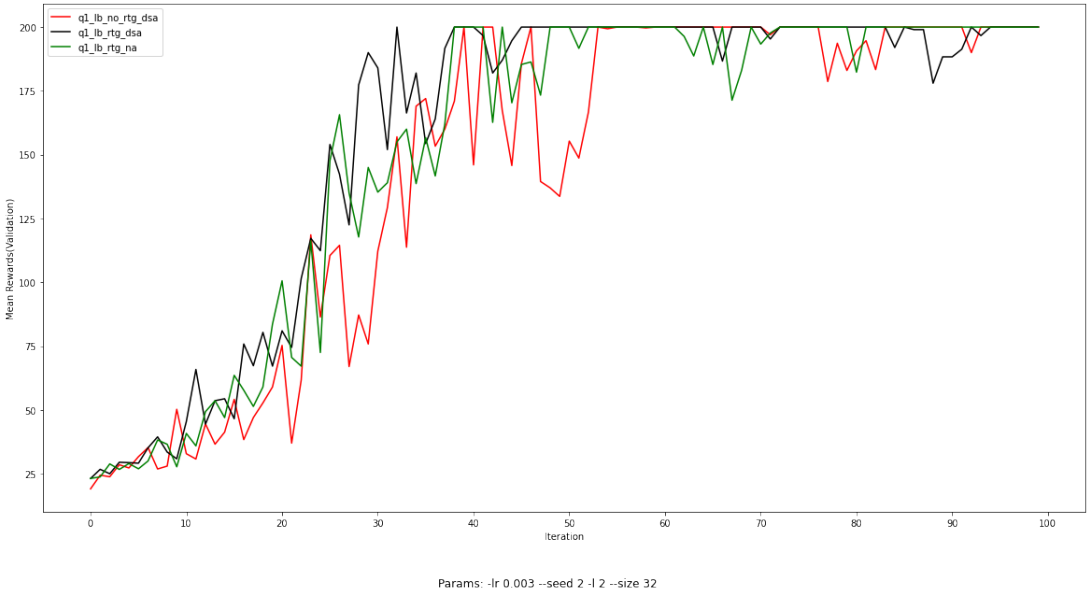
January 2023

1 Small-Scale Experiments

1.1 Experiment 1

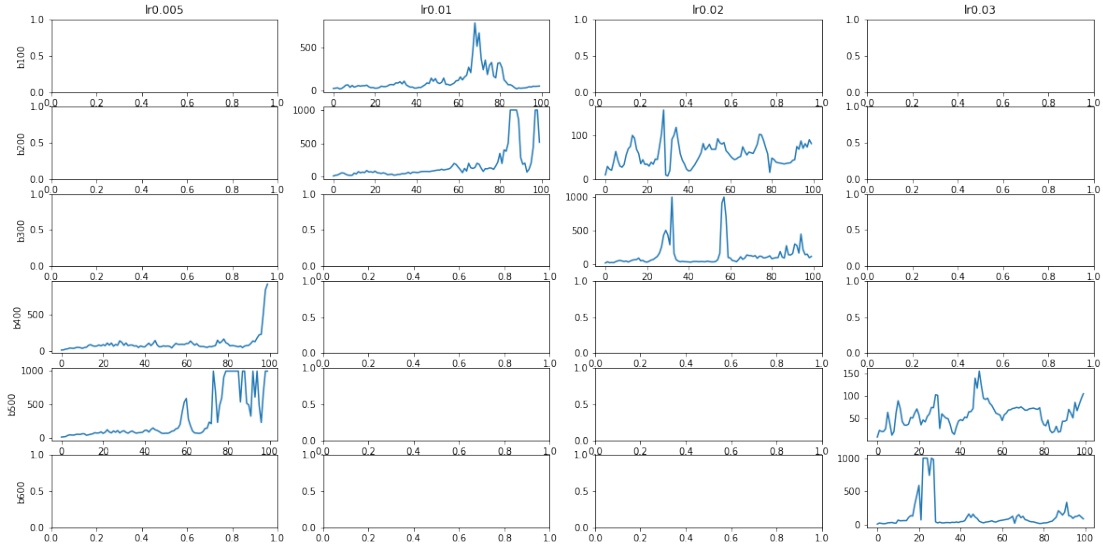


Reward-to-go estimator performs better. Advantage standardization difference is not statistically significant.



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 largest learning rate and smallest batch size that fails to reach 1000. In each column the second chart indicates largest learning rate and smallest batch size that succeeds in reaching 1000.