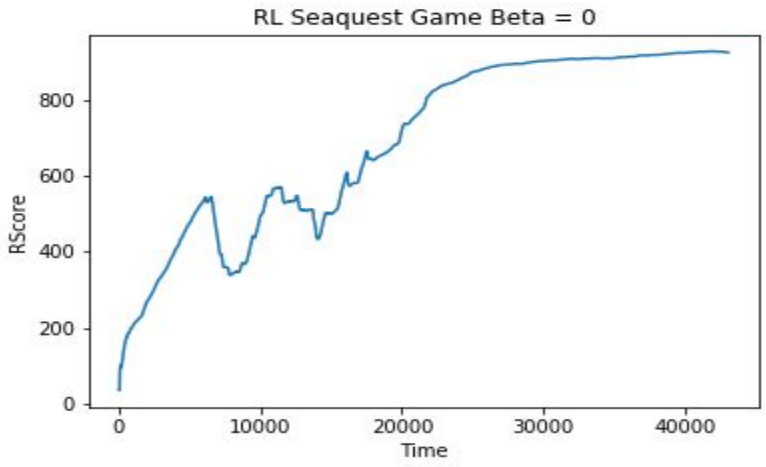
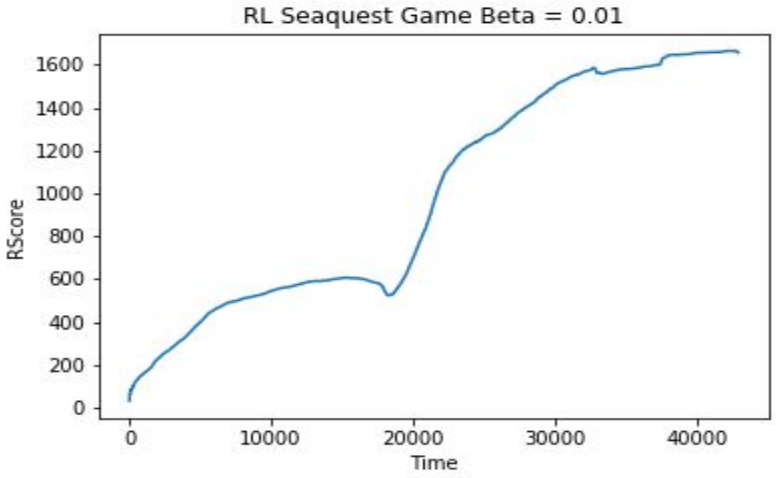
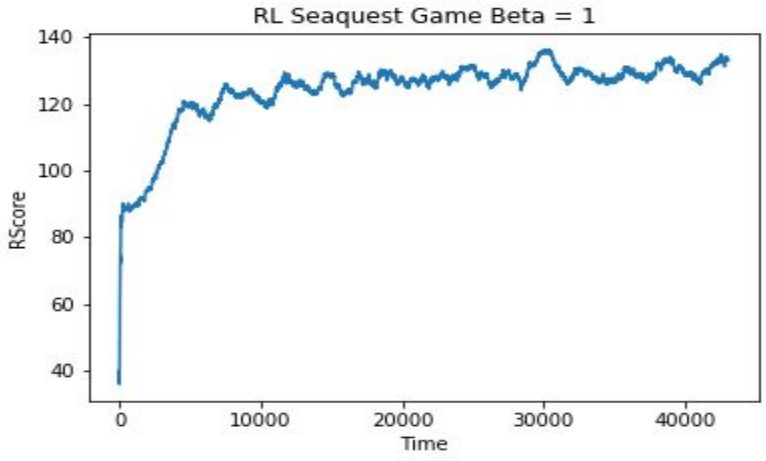
HW6 Reinforcement Learning Report:

[PART-1] Sensitivity to β - Try the following 3 values of β on the game Seaquest - {0, 1e-2, 1}. Report the learning curve, i.e. plot of average return (RScore from GA3C) vs. time. Explain what you see and try to justify.

The plots of Average Return (RScore from GA3C) vs. time for the Seaquest Game are shown below:



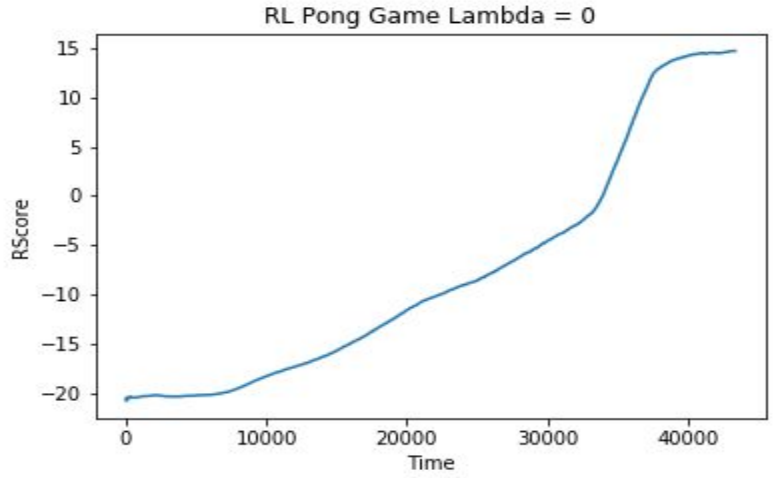


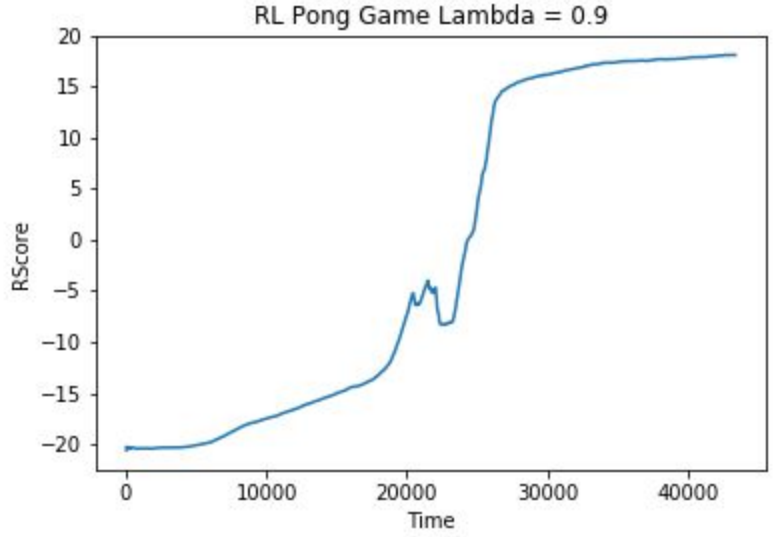


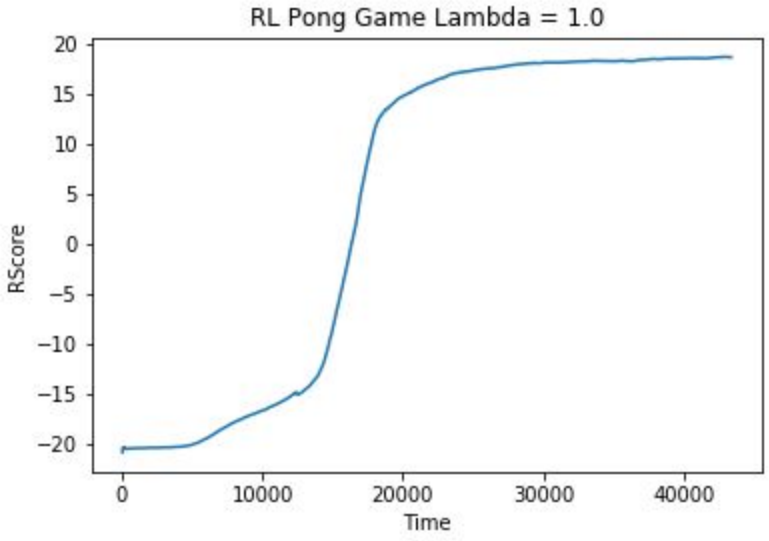
β is the weight on the entropy term, and therefore controls the amount of exploration the policy performs. Thus in the last graph with Beta = 1, this means that it is always exploring randomly even if it finds a good action. As a result it never approaches any sort of local or global maximum and performs poorly. In graph 1, with beta = 0, there is no exploration which means that it settles on a local minimum.

Finally with Beta = 0.01 there is a bit of exploration and this leads to the best performance and highest RScore as it does not get stuck at a local maximum peak.

• [PART-2] Bias-Variance Trade-off - In this part, you'll add GAE to the code. ProcessAgent.py has a function accumulate rewards which calculates the advantage estimate. The current code uses the high variance estimator (λ = 1). Modify this to calculate advantage using GAE, parameterized by λ. Run the code on the game Pong using the following 3 values of λ - {0, 0.9, 1}. Report the learning curve. Explain what you see and try to justify







In these graphs, for the game of Pong, the modified parameter is Lambda which controls the bias and variance in the learning. $\lambda = 1$ has high variance and low bias so it performs relatively well in terms of RScore and Accuracy even though it has a good amount of variance. The first figure of $\lambda = 0$, has very low variance because it is using mostly all estimations of V(S,t), but since it's only estimations and not using the actual actions it has very high bias and doesn't even get higher than 15 for the RScore. Finally, $\lambda = 0.9$ performs the best in terms of RScore because it has lower variance than the Lambda = 1 estimator and slight bias as well.