PA2 Report

Part 1. Compute Return

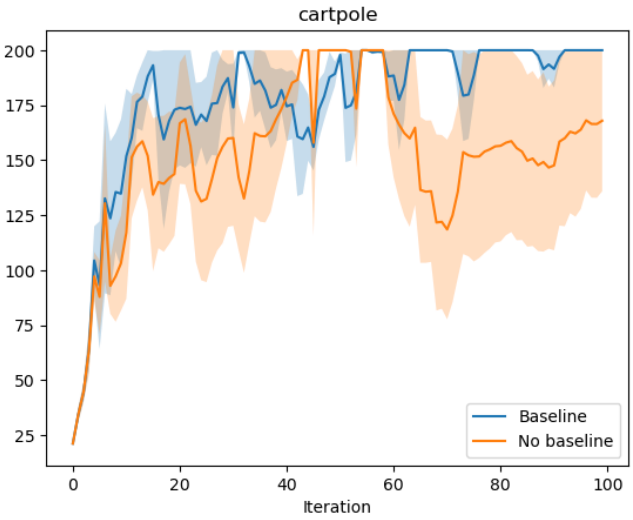
$$G_t = \sum_{t'=t}^{H} \gamma^{t'-t} r_{t'} \rightarrow G_t = \gamma G_{t-1} + r_t$$

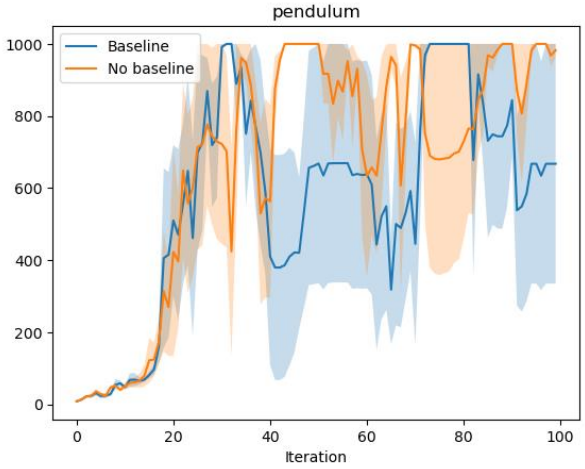
$For\ i\ in\ range(H)$:

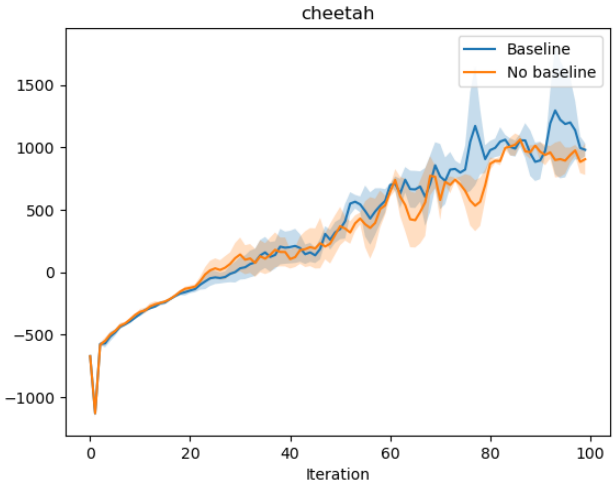
1. $calculate\ new\ return\ r_t$;
2. $update\ expected\ return\ G_t = \gamma G_{t-1} + r_t$

Based on that, the time complexity of computing return is $O(H)$.

Part 2. Reward Plots Under Three Environments







I chose seeds 1, 3, and 7 for Cartpole environment, seeds 1, 5, and 7 for Pendulum environment, and seeds 2, 3, 12 for Cheetah environment. From the plots shown above, implementing baseline under Cartpole and Cheetah environments achieves overall better performance, i.e., higher average reward score and more stable training process, than that without baseline. However, under Pendulum environment, implementing baseline to the network makes the training not stable which results in unignorable oscillations. It is uncommon, since the baseline can reduce variance for the estimator which makes the model more stable and achieves high average score of reward. I think it is caused by the specific seeds I chose for implementing.