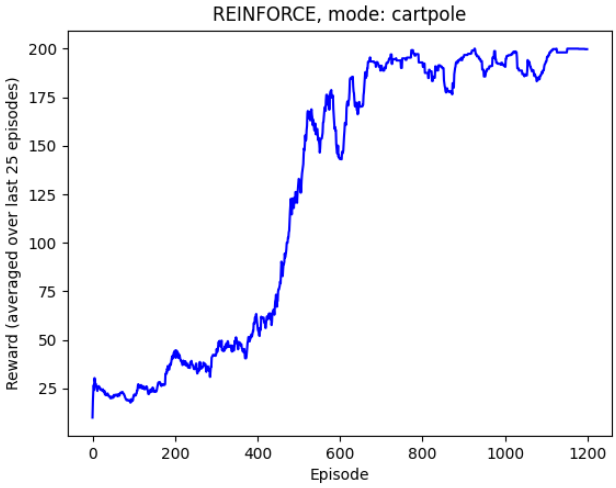
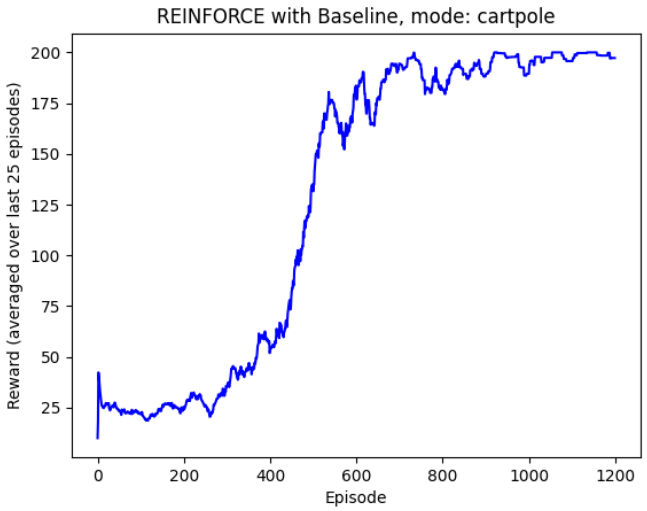
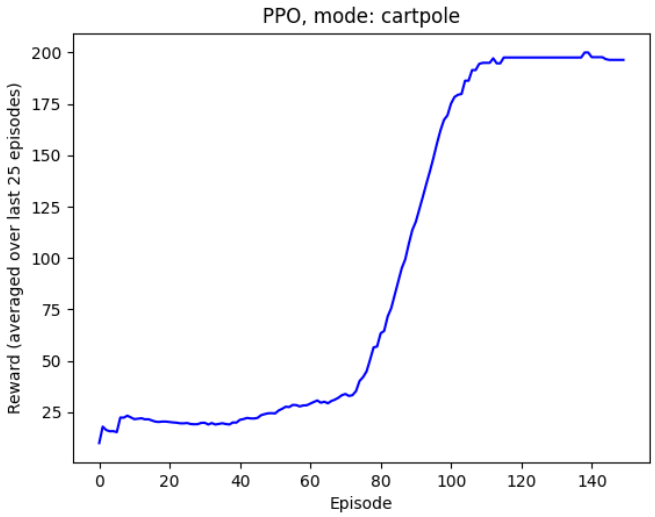
CS885 FA22: A2 PART 2







In the left, we show the produced graphs showing performance of REINFORCE, REINFORCE with BASELINE, and PPO on the cartpole problem. We observe that REINFORCE with BASELINE starts converging slightly earlier than REINFORCE with less variability.

For reference, we are showing 1200 episodes for REINFORCE vs REINFORCE with baseline. We see that cumulative rewards over the last 25 episodes are a lot smoother for REINFORCE with Baseline, and it achieves faster and stable convergence.

This is because the baseline used in REINFORCE with BASELINE is approximated value function. This approximation is based only on the current state, but not the actions. We know that state-action value function $Q^{\pi}(s,a)$ consists of value of executing action a, and value of executing policy π after executing action a. In the advantage function we remove the baseline which results in removing the value of executing policy π after executing action a from the Q function. This allows us to focus only on the immediate action at each step, removing the stochasticity associated with executing policy π after the action. This, results in less variance in reinforce with Baseline.

Compared to REINFORCE variants, PPO converges a lot earlier and faster. Setting learning rate α in Gradient update for policy is difficult, as small learning rate results in slow but reliable convergence, and large learning rate results in fast, but unreliable convergence. To solve this issue, we define a trust region around our current policy π, within which the update in policy is smooth. We are using small, predictable changes in the value function V with the use of trust region, compared to the unreliability of the learning rate. So, in trust region technique – we try to maximize our policy value in such a way that it maximizes but do not cross the advantage function value. What this results in is achieving the same effect as REINFORCE-BASELINE in focusing on only the immediate action in each step, while avoiding the erraticness associated with learning rate α. This results in a faster convergence compared to REINFORCE with BASELINE.

Arefin Shimon, Shaikh Shawon 20942174 ssarefin@uwaterloo.ca