Policy Gradient with Baseline

Sungsoo Lim

November 5, 2020

1 Baseline calculation

For both CartPole and Pong games, a second DNN was trained to compute the value estimation for the given state. Then, the value estimations were used to first calculate the losses for the policy estimation DNN and value estimation DNN, where the loss for the policy DNN was the softmax cross entropy loss minus the value estimation, and the loss for the value DNN was the mean squared error between the reward and the value estimation. Discounted rewards were also calculated with the rewards and rewards with baseline of the value estimations. Gradients of the policy network were updated with the discounted rewards and the those of the value network were updated with the discounted rewards with baseline.

2 CartPole-v0

Batch size of 10 was used for the simulation.

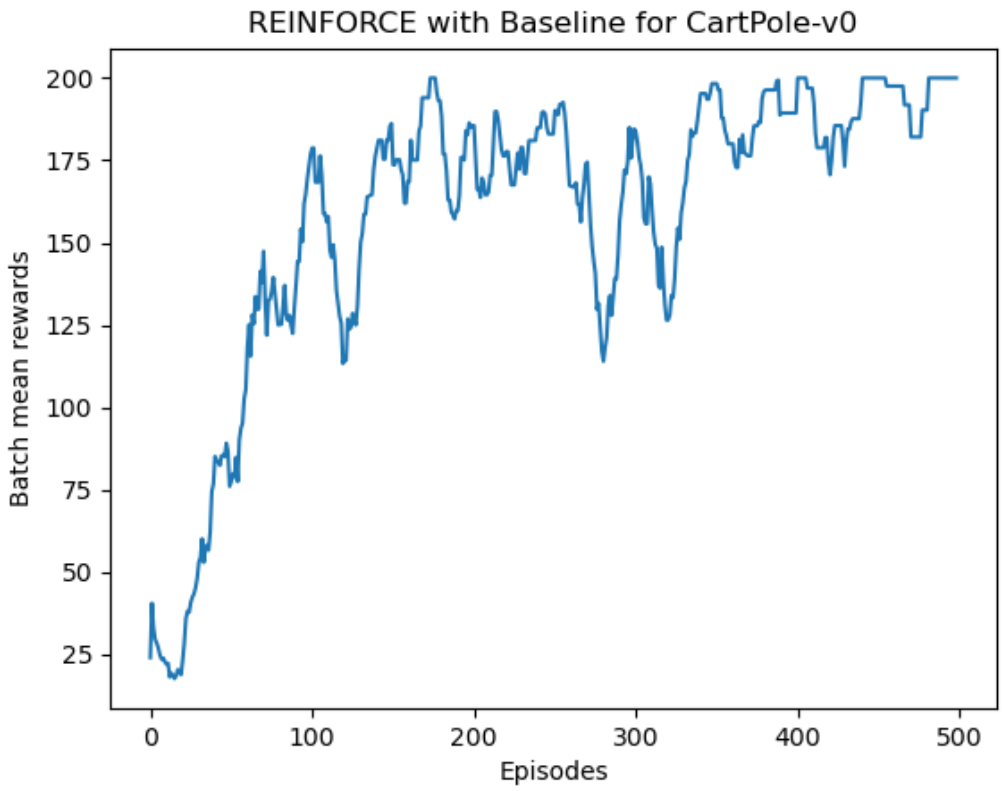

Figure 1: Batch-mean total rewards vs. episode numbers for the REINFORCE algorithm with baseline for CartPole-v0.

3 Pong-v0

Batch size of 10 was used for the simulation. For 1000 episodes, the algorithm was not yet learning. The learning rates for the policy and value DNNs can also be experimentally validated for more efficient training.

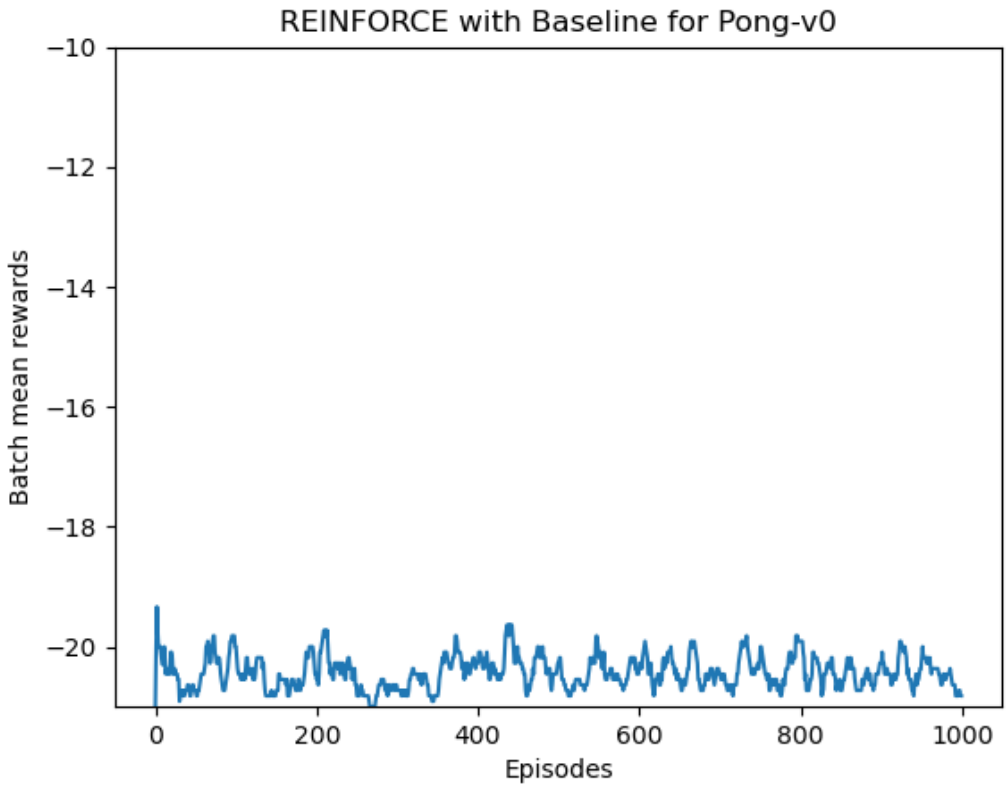

Figure 2: Batch-mean total rewards vs. episode numbers for the REINFORCE algorithm with baseline for Pong-v0.