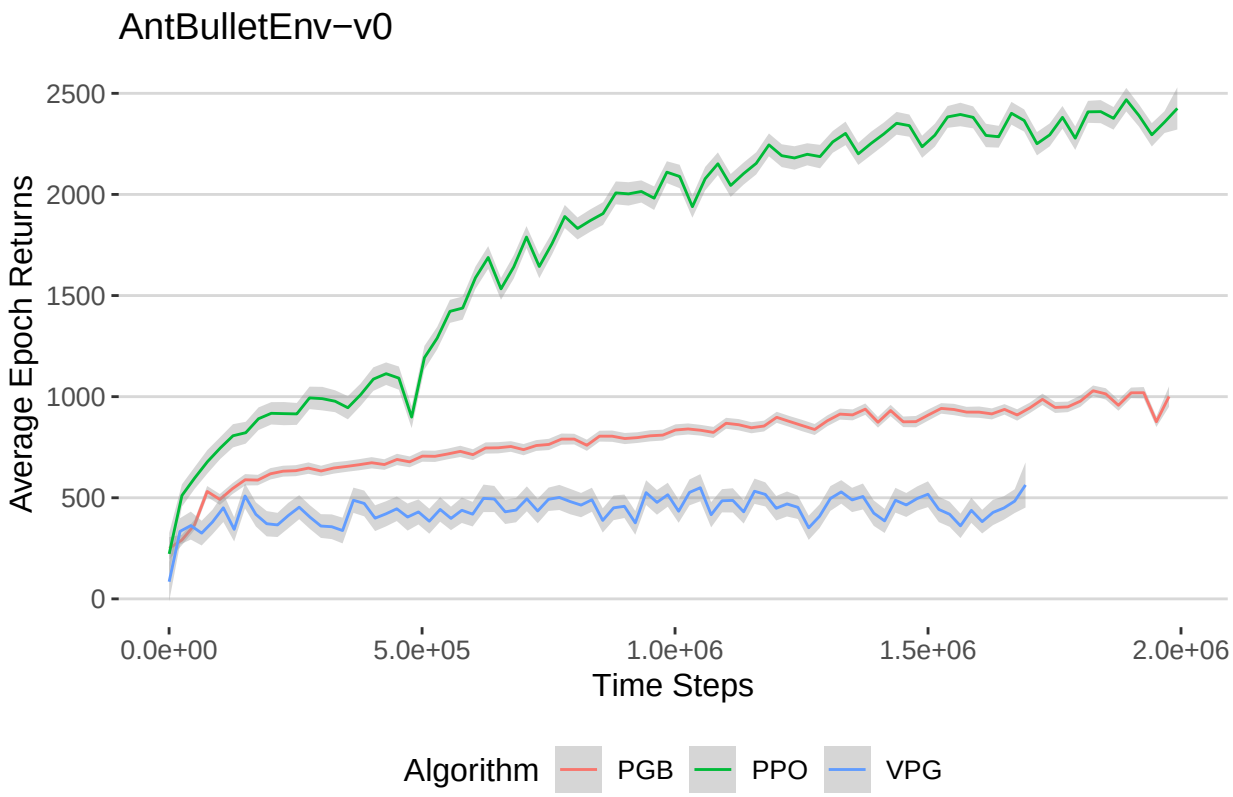
Homework 1 Report

Sang Doan

10/14/2021

```
selected_cols <- c("time_steps", "AverageEpReturns")
dat <- rbind(
  vpg[, ..selected_cols][, `:=` (algo = "VPG")],
  pgb[, ..selected_cols][, `:=` (algo = "PGB")],
  ppo[, ..selected_cols][, `:=` (algo = "PPO")]
)

plot_algos(dat)
```



Epoch returns are averaged every 10 loops.

Overall, out of the three algorithms, PPO performed the best regarding optimal outcomes, running time and computational efficiency. PGB yielded the lowest variance, but even after roughly 2 million steps, it couldn't go beyond the 1,000 mark very far. VPG was the worst performer, with high variance and low scores overall.

Concerning training time, all three were comparable. On the same machine, VPG took 50 minutes to complete 1.7 million steps. PPO took 1.2 hours and PGB 1.4 hours to complete 2 million steps. But we can

see that PPO started to converge much earlier than the other two; it reached a score of 2,000 after only 1 million steps.

Two observations— First, during training, I noticed that PGO and especially are, in general, sensitive to learning rates or hyperparameter tuning, while PPO less so. Second, an interesting thing from the graph is that throughout the training, the PPO agent appeared to make several mistakes (the deep dips) but quickly recover each time. This might point to PPO's high recoverability from mistakes.