Reinforcement Learning Homework 04

Qiu Yihang

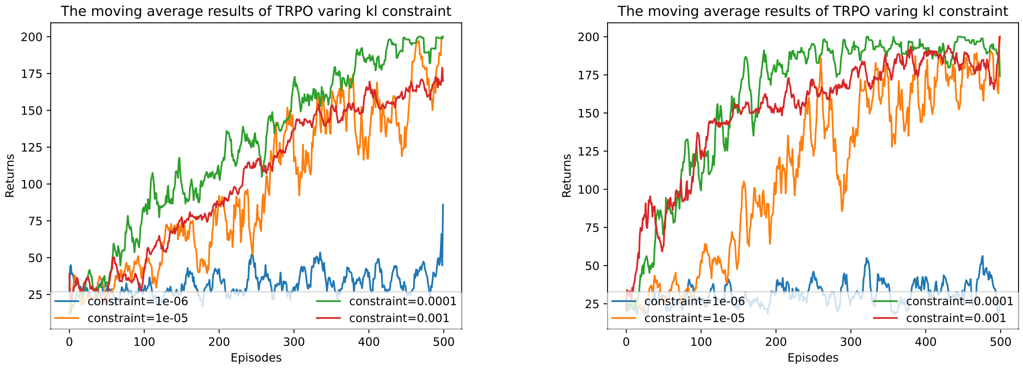
May 2023

1 Implementation of TRPO and PPO algorithms

The source code is under directory `code`. The results and analyses are as follows.

1.1 TRPO with different trust region constraints

The performance of TRPO algorithms with different trust region constraints δ is as follows.



(a) random seed=14985

(b) random seed=20210530

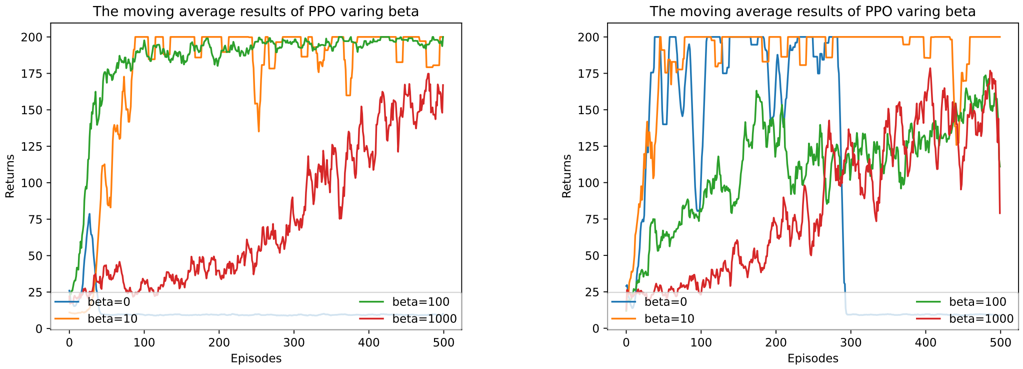
Figure 1: The performance of TRPO algorithms with different values of trust region constraints δ in environment $\texttt{CartPole} - \texttt{v0}$.

It is plain to see that the trust region constraints with too small or too big values will lead to a decrease in performance, especially small δs. Small δs will keep the agent stuck in a certain policy and prevent it from exploring to a better one. On the other hand, big δs will make the agent act out of the relatively trusted region sometimes and thus lead to a minor decrease in performance.

The best performance is achieved when $\delta = 0.0001$.

1.2 PPO with different fixed penalty coefficients

The performance of PPO algorithm with different fixed penalty coefficients β is as follows.



(a) random seed=14985

(b) random seed=20210530

Figure 2: The performance of PPO algorithms with different values of fixed penalty coefficients β in environment $\texttt{CartPole} - \texttt{v0}$.

It can be seen that penalty coefficients with too small or too big values will cause the performance to decrease. The reason is similar to that of TRPO algorithms. Small βs lead to greater tolerance on the KL divergence and the agent is more likely to fall into untrusted regions. Large βs lead to more strict constraints on the KL divergence and the agent is more likely to be stuck in a certain policy, which hurts the exploration and decreases the performance.

The best performance is achieved when $\beta = 10$.

1.3 The Similarity Between δ of TRPO and β of PPO

The impacts of the two parameters are similar. Both δ in TRPO and β in PPO adjusts how different the distribution of the new policy can be away from the old one. They help to explore policies in a relatively trusted region and balance the exploration and exploitation.