1. (Implementation of the TRPO and PPO algorithms) You are required to implement the TRPO and PPO algorithms for the gym CartPole-v0 environment. In this problem, a pole is attached by an unactuated joint to a cart, which moves along a frictionless track. The pendulum is placed upright on the cart and the goal is to balance the pole by applying forces in the left and right direction on the cart.

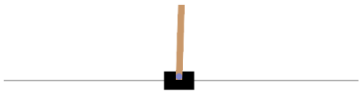

Figure 1: The Cart Pole environment.

In this environment, the observation is a `ndarray` with shape `(4,)` with the values corresponding to the following positions and velocities:

Num	Observation	Min	Max
0	Cart Position	-4.8	4.8
1	Cart Velocity	-Inf	Inf
2	Pole Angle	~ -0.418 rad (-24°)	~ 0.418 rad (24°)
3	Pole Angular Velocity	-Inf	Inf

Figure 2: Observation space.

The action is a `ndarray` with shape `(1,)` which can take values $0, 1$ indicating the direction of the fixed force the cart is pushed with:

- 0: push the cart to the left;
- 1: push the cart to the right.

For the TRPO algorithm, you are required to implement:

(a) the `line_search()` function to find the best hyper-parameter for actor update (you may refer to [1, Appendix C]);
(b) the `policy_learn()` function to update actor policy (you may refer to [1, Section 6]);
(c) the `update()` function to update the parameter of actor and critic (you may refer to [1, Section 6]).

For the PPO algorithm, you are required to implement:

(a) the `update()` function to update the parameter of actor and critic (you are required to implement the penalty version of PPO with fixed penalty coefficient β, and you can refer to [2, Section 4]).

Run your codes to show the performances of TRPO and PPO in the above environment (please include the figures of experimental results in your report). And answer the following questions:

(a) Compare the performance of TRPO algorithms with different values of trust region constraints, i.e., δ in [1]. Discuss the reason for this difference.
(b) Compare the performance of PPO algorithms with different values of penalty coefficients in the KL-divergence term, i.e., β in [2]. Discuss the reason for this difference.
(c) Are the impacts of parameter δ on TRPO algorithms and the impacts of parameter β on PPO algorithms similar or not? Discuss the reason.

References

[1] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[2] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.