|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | Network | Policy | Action Space | Operator |
| DQN family | Critic | Off-policy | Discrete | Q-value |
| DDPG | Actor-Critic | Off-policy | Continuous | Q-value/Policy Gradient |
| PPO | Actor-Critic | On-policy | Discrete/Continuous | Policy Gradient |

Empirical Study

Since the DDPG and PPO are both designed to solve problem in continuous action space, we’ve chosen the “LunarLanderContinuous-v2” environment to perform the empirical study. Same as the discrete action space version of Lunar Lander game, the objective is to land lunar lander in the landing pad with the possible highest score. However, the input action is not changing from 1 of 4 discrete actions; instead, it is into a two values vector, the first value controls main engine power, and the second controls left and right engines, ranging from -1 to +1. Both algorithms are executed in Google Colab, DDPG is implemented with Pytorch V1.81 and PPO with Tensorflow V2. Plus , to keep all tests possibly equal, we set the environment seed to 123 for all runs.

# **DDPG**

The most important difference between DQN and DDPG is that the latter adopts the Actor-Critic method: Actor Network outputs action and Critic Network estimates action value (Q-value). Critic Network and Actor Network work together as following:

1. Actor Network outputs action based on observation received from environment
2. Environment returns rewards of received action also next observation
3. Update Critic Network based on rewards, then update Actor Network

Actor-Critic structure becomes popular in Reinforcement Learning and is adopted in many most advanced algorithms, such as Proximal Policy Optimization (PPO) and Soft Actor-Critic (SAC).

## Setup

DDPG has many hyperparameters, most important are followings:

* **Gamma**: this parameter affects the weight of rewards from next steps. We expect that the agent considers not only the reward from next steps but also from longer future. The smaller this value, the higher the weight of next few steps. But also, the longer future we take into consideration, the lower the quality of prediction. The balance between them needs to be made, so we set this value to 0.99.
* **Tau**: DDPG uses “soft update” slowly in every step instead of “hard update” after every X steps as in DQN. This value defines how “soft” the update will be. The smaller the value, the “softer” the update. We set it to 0.001.
* **Memory size**: DDPG stores previous experiences and sample from them to train networks. Once memory size is exceeded, we will replace oldest experiences with new ones. We set it to 2 \*\* 17.
* **Batch size**: How many samples we will use to train the model. The lower this value, the faster the training speed, but it also possible to reduce the convergence speed (new more episodes to converge). We set it to 128.

Besides of those DDPG related parameters, we also used following network parameters:

* Actor network:
  + 2-layers Neural Network
    - 512 hidden units in 1st layer
    - 256 hidden units in 2nd layer
  + Learning rate: 5e-5
* Critic network:
  + 2-layers Neural Network for state input
    - 512 hidden units in 1st layer
    - 256 hidden units in 2nd layer
  + 1-layer Neural Network for action input
    - 128 hidden units
  + 1-layer Neural Network with concatenated outputs from state part and action part as input
    - 128 hidden units
  + Learning rate: 5e-4

In noise for exploration part, we used 2 types of noise:

* Ornstein-Uhlenbeck (OU) Noise mentioned in DDPG original work Lillicrap et al. (2015)
* Gaussian Noise

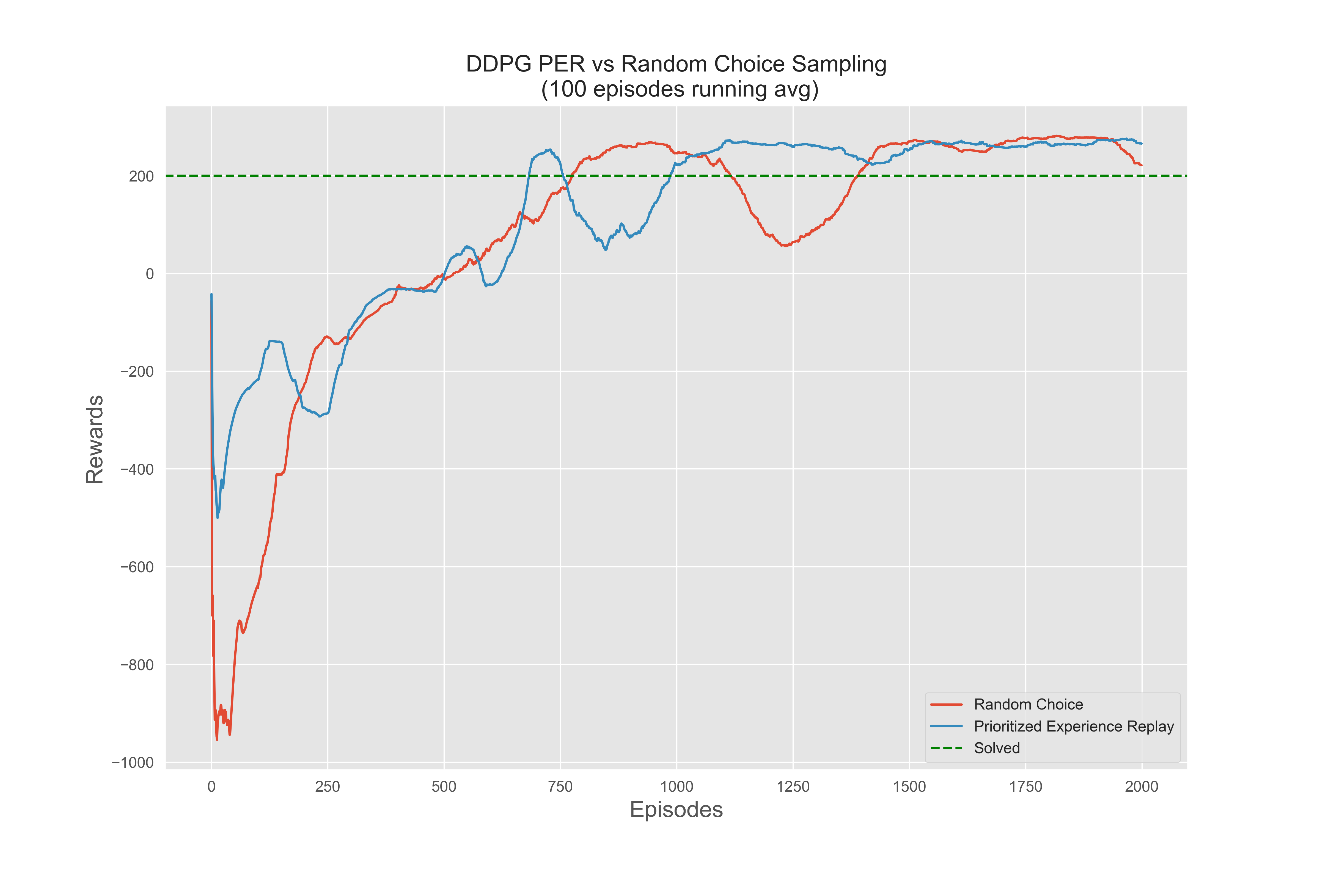
For both noises, the mean is set to 0 and standard deviation is set to 0.2.

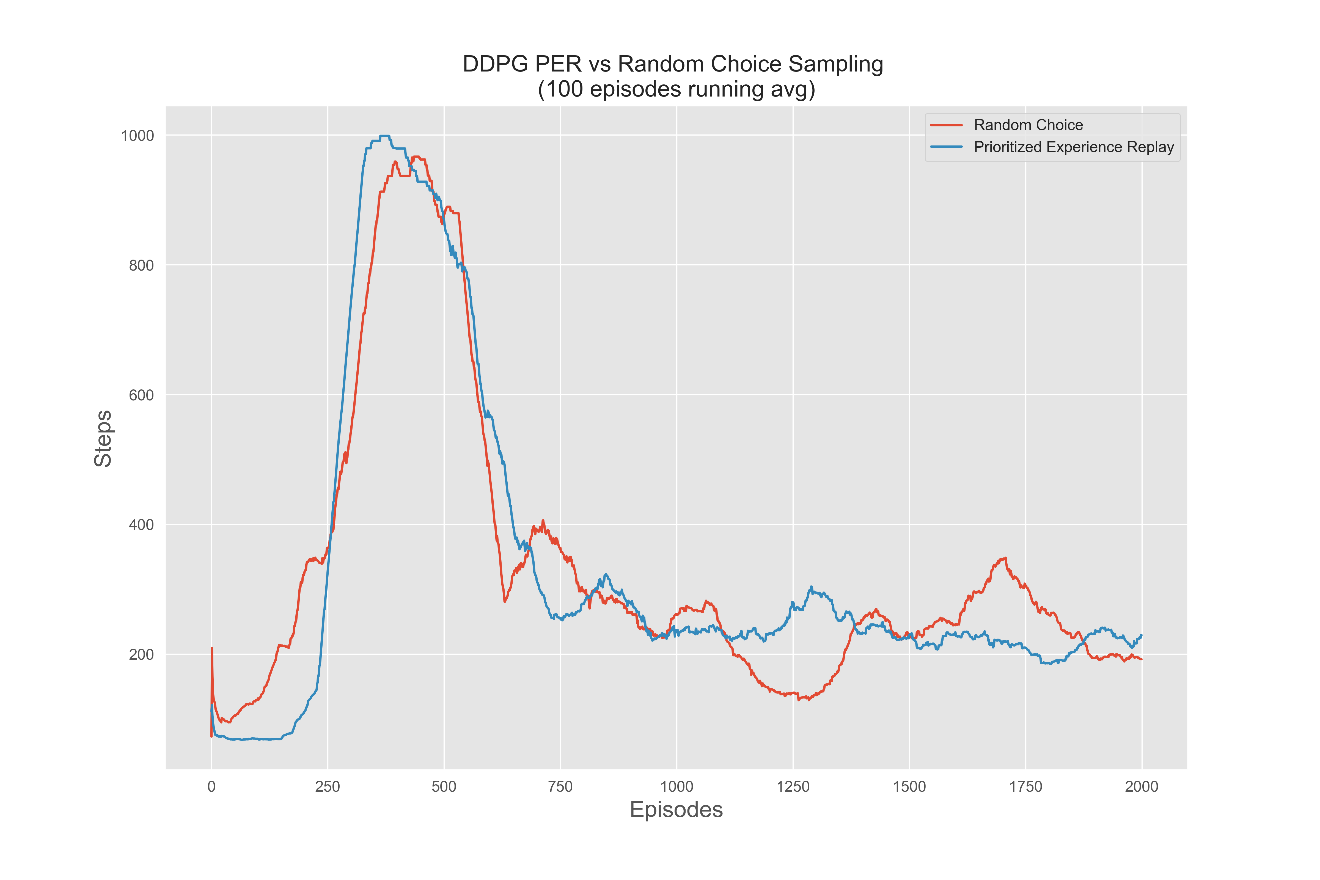
## Results

One major objective of this empirical study is to understand the prioritized experience replay (PER) performance. In the original DDPG algorithm, the sampling method from stored experience is uniformed random sampling, which means every experience in memory has the same probability to be selected. But as common sense tells us, not all the experiences are the same important for learning; if we are able to find the most helpful experiences, it would be helpful to reduce learning time. PER used TD error to measure the importance of experiences and prioritize them while sampling. According to authors (T.Schaul et al. 2016), the PER is able to speed up the convergence and achieve a better performance compared to DQN using uniform sampling experience replay. Since the Critic part of DDPG is almost the same as DQN, can PER improve the convergence speed and performance in DDPG?

As the rewards figure shows, PER did quickly increase rewards in first 200 episodes, but since the 250th episode, those two methods kept almost the same pace in scoring. Although PER version reached Solved (*scored more than 200 points in an episode is considered as problem solved*) line earlier than Random Sample version, it’s not always the case in later runs. So, we cannot prove that PER brings better performance.

Since DDPG updates networks every step, using fewer steps could reduce the training time and speed up convergence. While checking average steps chart, we can find that, in the first 250 rounds, PER version did use less steps than Random Sample version. But for the whole training period, Random Sample version used only 0.2% fewer steps than PER version. As both models achieved stable performance after 1500 episodes, if we consider only first 1500 episodes, Random Sample version used 3.7% fewer steps than the PER version. In this case, we cannot find enough evidence to support the assumption that PER would bring better performance and convergence speed to the DDPG algorithm.

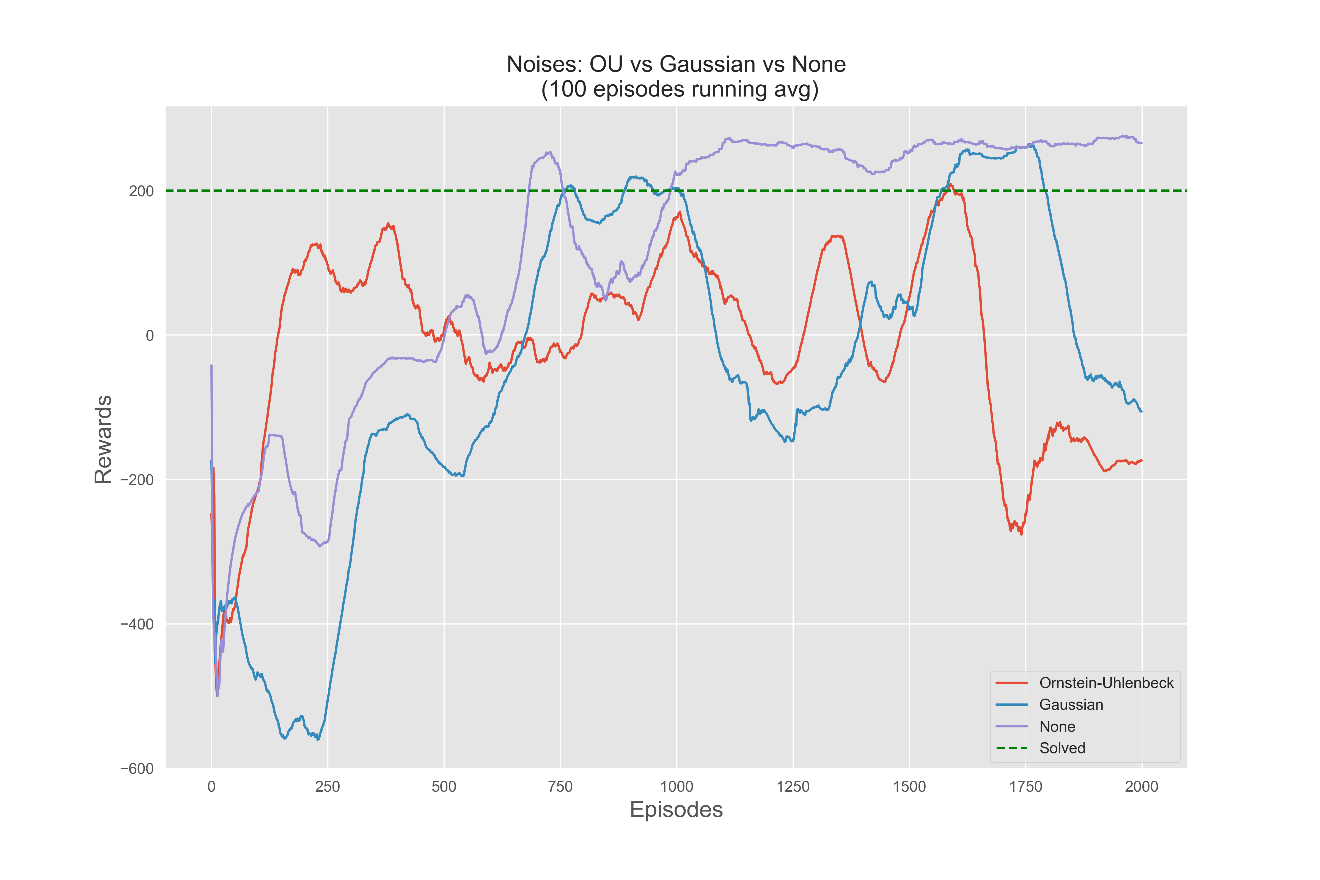




Another objective is to test the performance of different action noises. As mentioned in DDPG original paper, Lillicrap et al. (2015) use OU Noise in exploration. Is this a better choice than Gaussian noise? What if we don’t use noise in exploration at all?

Described in DDPG paper, Lillicrap et al. (2015) adopt OU noise to generate temporally correlated exploration, which has higher exploration efficiency in physical control problems with inertia. One major difference between OU noise and Gaussian noise is that the absolute difference between each continuous two steps in OU noise is not as great as in Gaussian noise. Also, continuous steps have trend to follow the direction of momentum. Landing control in our environment is exactly the system described: each action will generate feedback with inertia.

From Noise graph, we can see that all noise versions could not converge, although they all reached the solved line, but the performance is not stable at all. Among all three versions, the one that has the best performance does not use noise. And OU noise has even worse performance than Gaussian noise. After multiple rounds of tests, we found that, without noise, each time the model can achieve convergence and get a stable performance. On the contrary, neither OU nor Gaussian noise can get similar results with our configuration (mu=0, sigma=0.2). Due to the usage limit of Google Colab, we were not able to run more tests with different parameters, which could be an interesting subject in future research.



**PPO**

Since Schulman et al. (2017), PPO has been using as default Deep RL algorithm by OpenAI, and as the 1st choice for many RL solutions because of its simplicity, powerful performance, quick training speed, and versatility.

Since the core of policy gradient is to increase the probability of “good” samples and reduce the probability of “bad” samples. So, a common problem that all policy gradient (PG) algorithms are facing is the instability of training caused by the sparseness of rewards in experience and sampling. Imaging for a sample which occasionally having a very high reward its probability will be increased rapidly, comparably, for one having a very low reward, its probability will be decreased rapidly, this situation generates instability of training.

To increase the stability, different PG algorithms use different methods to reduce the frequency of updates or to limit changes between updates, for example, the “soft update” we saw in DDPG. The method used in TRPO is to do an optimization constrained by KL divergence through conjugate gradient, which limits changes between updates. But this method requires to calculate second derivative of KL divergence, which increases the complexity and could use many resources. PPO improved this method by moving the constraints into loss function, which speeds up computation and reduces computation needs. The optimization in PPO can be described as following:

* If the new policy is improving towards the right direction, the algorithm will stop optimizing the new policy once over a certain threshold. But if the new policy is getting worse, the algorithm will keep optimizing it.
* If the new state value is closer to target than old one and over a certain threshold, the algorithm will stop optimizing the state function. But if contrary, the algorithm will keep optimizing it.

## Setup

Compared to DDPG, PPO has fewer hyperparameters, most important are the followings:

* **Lambda advantage**: Similar to Gamma in DDPG, this parameter affects the weight of rewards from next steps. Smaller Lambda advantage brings more weights to next few steps. We set this value to 0.99 as in DDPG.
* **Clip epsilon**: this parameter limits the range of changes. As described in introduction, PPO uses this parameter as the threshold to limit amplitude of updates. We set this value to 0.2.
* **Number of updates**: this parameter controls the number of updates each episode, we set this value to 10.

By the way, although PPO is an on-policy algorithm, it’s still able to learn from experiences from previous episodes. But to compare with off-policy algorithms, we forced PPO only to learn from current episode, which means we don’t maintain an experience memory.

Besides those PPO related parameters, we also used following network parameters:

* Actor network:
  + 2-layers Neural Network
    - 512 hidden units in 1st layer
    - 512 hidden units in 2nd layer
  + Learning rate: 1e-5
* Critic network:
  + 2-layers Neural Network
    - 512 hidden units in 1st layer
    - 512 hidden units in 2nd layer
  + Learning rate: 1e-3

## Results

The major objective of this empirical study is to compare PPO and DDPG, from performance, execution time and other prospects.

First, on comparing the execution time, we found PPO is much quicker than DDPG. In general, it takes PPO 30-40 mins to complete 2000 episodes, but DDPG needs at least two and half hours to finish the same run.

Second, from the comparison graph, we can see that PPO’s rewards curve increases much more smoothly than DDPG runs. Although DDPGs achieved higher average scores in the last 1000 episodes, it’s not always the case in different runs. Also, PPO started to increase the average rewards at the very beginning compared to DDPG algorithms and didn’t touch a very negative reward as DDPGs did during the entire run, which reveals the stability of the algorithm.

Third, although the Lunar Lander game is limited the maximum steps to 1000 per episode by default, we think this value is too high. Higher steps limit may encourage agent to explore more, we don’t want a dawdling spaceship to waste much time in exploration. Also, we noticed that, for both PPO and DDPG, the number of used steps per episode increases rapidly in the first half of training, which means the agent was trying to stay in the air to avoid crash penalty, but by the end of training, this number always reduced to less than 300 steps, which means the agent learned that quickly landing on the correct zone is a better strategy.

Furthermore, we designed test runs with different maximum steps limit. From the figure, we can see that the maximum 300 steps run shows the least velocity among all tests, but the other two runs reached “resolved” line earlier than it. The fact indicates that limiting maximum steps is indeed helpful to achieve better performance. Also, limiting maximum steps makes each episode taking less time to run. In general, a maximum 300 steps version spent only 55% of time needed to run default maximum steps (1000) version. Both advantages make the maximum steps limitation an attractive choice.

By curiosity, we also ran same tests with DDPG and got similar results: lower maximum steps version needs more episodes to score higher rewards, but it shows less velocity and spends less time to finish training. But DDPG with 300 maximum steps was not able to reach the same level of rewards as the other two (500 and 1000) versions in test. The reason could be the instability of algorithm. Due to the limitation in execution environment, we were not able to run more tests, it could be a subject for future research.

