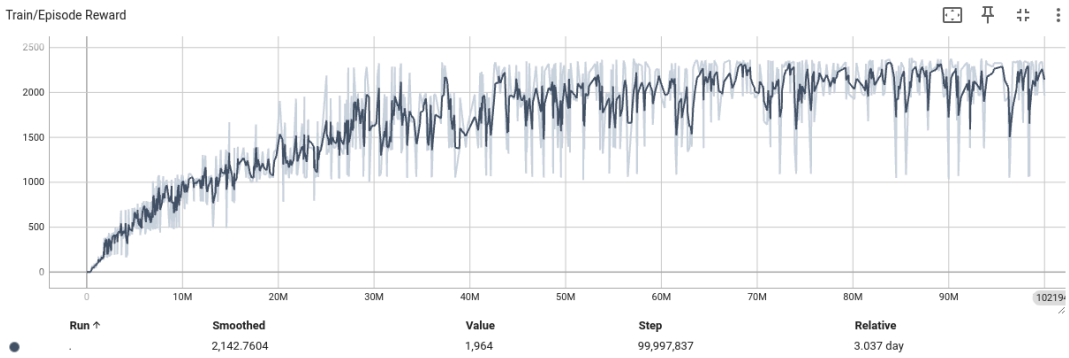
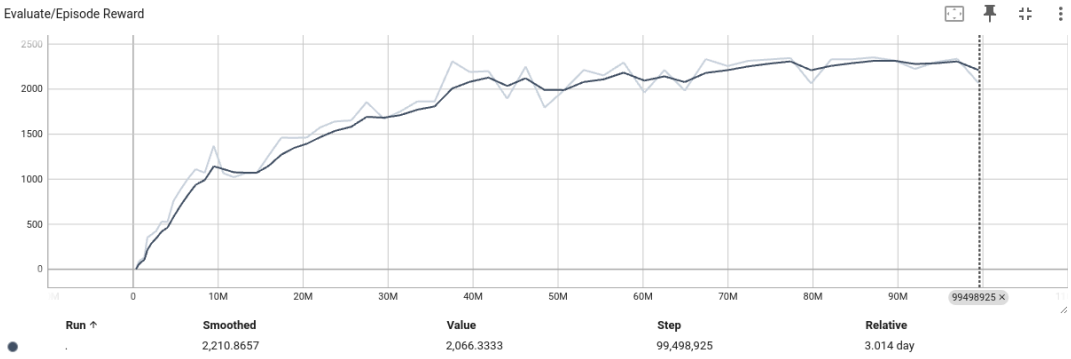
Lab3: PPO

Student Name: 歐庭維
Student ID: 312605001

1. Solve Enduro-V5 using PPO
    - Training Curve



    - Evaluate Curve



    - Test in 5 Games

```
episode 1 reward: 2374.0
episode 2 reward: 2383.0
episode 3 reward: 2347.0
episode 4 reward: 2387.0
episode 5 reward: 2373.0
average score: 2372.8
```

2. Question
    a. PPO is an on-policy or an off-policy algorithm? Why?

    Ans:

    The difference between on-policy and off-policy lies in whether the agent interacting with the environment is the same as the one being trained. Although PPO is typically an on-policy algorithm because it primarily uses sample data generated by the current policy for updates, we have introduced a replay buffer in our PPO training process to reuse past sampled data. This enables our algorithm to leverage previous experience, which is an off-policy characteristic. As a result, our PPO implementation combines the advantages of both on-policy and off-policy approaches, which allow the model to reuse the data efficiently and reduce the need for new data.

b. Explain how PPO ensures that policy updates at each step are not too large to avoid destabilization.

Ans:

The training objective of PPO is to maximize the objective function to find the better policy while ensuring that the difference between the new and old policies does not become too large, thus avoiding policy collapse and improving training stability. In the following objective function, the latter term represents the KL divergence between the new and old policies, which indicates the difference between them. However, calculating the KL divergence is complex, so PPO2 was developed to approximate KL divergence using a simpler method, which is the algorithm we are using in this assignment.

$$J_{PPO}^{\theta^K}(\theta) = J^{\theta^K}(\theta) - \beta \mathrm{KL}(\theta, \theta^K)$$

In the following PPO 2 formula, the use of the min function along with the clip function ensures that policy updates remain within a certain range, with the size of this range defined by epsilon. When the new and old policies are close, meaning the ratio r(θ) is close to one, the clip operation does not take effect, and the loss function directly calculates the product of the ratio and the advantage function. When the ratio deviates from one and exceeds the clipping range, the clip mechanism is activated, restricting the ratio within the range defined by one minus epsilon to one plus epsilon. This prevents the advantage function from having an excessive impact and limits the update magnitude between the new and old policies.

$$J_{PPO2}^{\theta^K}(\theta) \approx \sum_{(s_t,a_t)} \min\left(\frac{p_\theta(a_t|s_t)}{p_{\theta^K}(a_t|s_t)} A^{\theta^K}(s_t,a_t), \mathrm{clip}\left(\frac{p_\theta(a_t|s_t)}{p_{\theta^K}(a_t|s_t)}, 1-\epsilon, 1+\epsilon\right) A^{\theta^K}(s_t,a_t)\right)$$

c. Why is GAE-lambda used to estimate advantages in PPO instead of just one-step advantages? How does it contribute to improving the policy learning process?

Ans1:

One-step advantages rely solely on the current reward and the next step's predicted value, making them susceptible to reward noise, which results in high variance. GAE-λ reduces variance by computing a weighted average of multi-step advantages. It helps PPO smooth the advantage estimates, reducing the impact of short-term fluctuation that can lead to policy instability, thereby stabilizing the policy update process.

Ans2:

As the time horizon increases, GAE-λ incorporates longer-term reward information, which helps smooth out fluctuations caused by short-term noise. Therefore, when λ is close to 1, the variance of the advantage estimation is lower, making the advantage values more stable and less susceptible to short-term reward fluctuation.

d. Please explain what the lambda parameter represents in GAE-lambda, and how adjusting the lambda parameter affects the training process and performance of PPO?

Ans:

In GAE-λ, λ is essentially a discount factor used to adjust the weighting of future rewards in advantage estimation. Adjusting the parameter λ balances the bias and variance in advantage estimation: setting λ=1 represents using longer time horizons to estimate advantage, which reduces variance but introduces bias; setting λ = 0 makes the behavior closer to one-step advantages, with low bias but high variance.