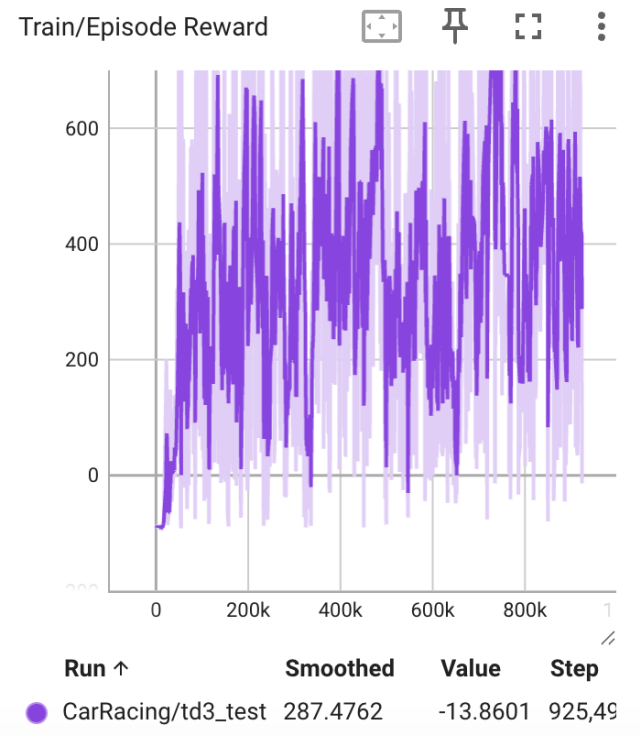
Lab 4: Twin Delayed DDPG (TD3)

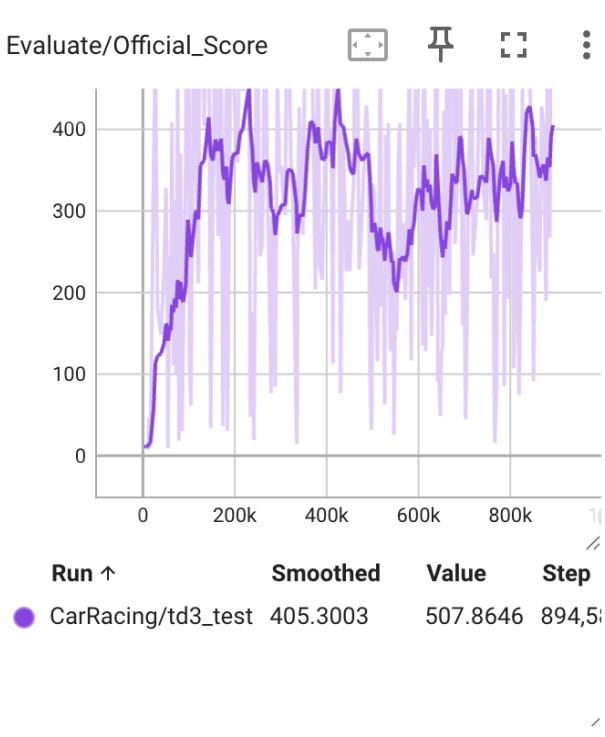
314553017蕭希敏

Result

> 所有的**Evaluate /Official_Score**是因為我Train /Episode Reward的test設為false，重新用所有.pth重新evaluate得到的數據

```
Episode: 1      Length: 186     Total reward: 156.22
Episode: 2      Length: 481     Total reward: 538.61
Episode: 3      Length: 337     Total reward: 333.08
Episode: 4      Length: 999     Total reward: 866.91
Episode: 5      Length: 999     Total reward: 850.00
Episode: 6      Length: 999     Total reward: 847.57
Episode: 7      Length: 999     Total reward: 872.13
Episode: 8      Length: 999     Total reward: 821.64
Episode: 9      Length: 518     Total reward: 445.37
Episode: 10     Length: 568     Total reward: 491.88
average score: 622.3398945917354
================================================
```

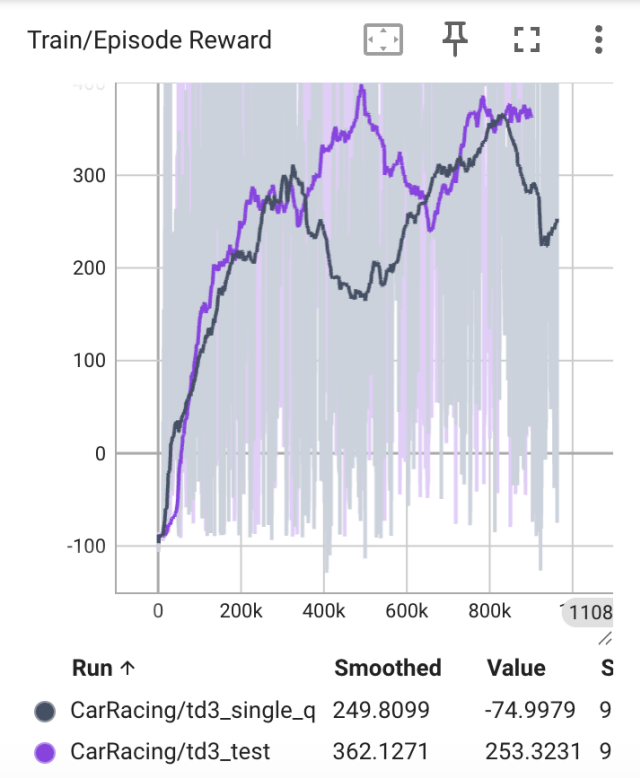


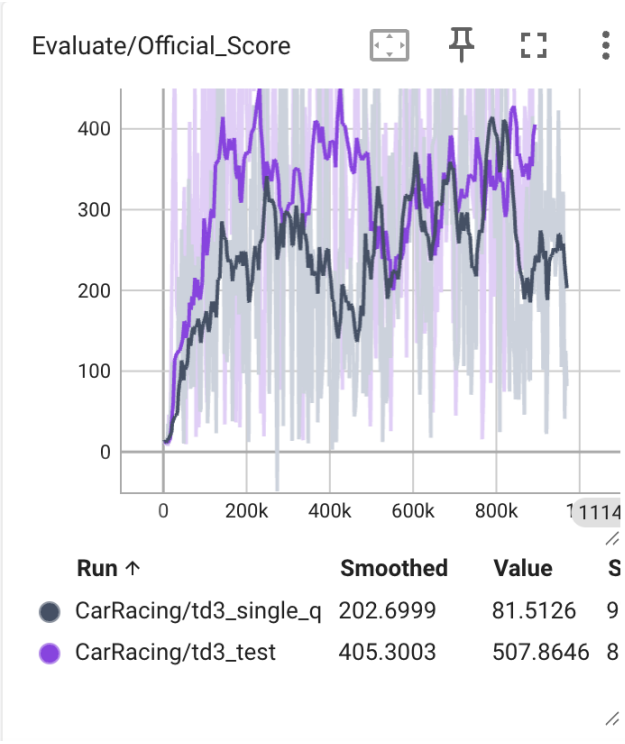


Discussion

1. **Screenshot of Tensorboard training curve and compare the performance of using twin Q-networks and single Q-networks in TD3, and explain.**

```
Episode: 1      Length: 429     Total reward: 590.12
Episode: 2      Length: 122     Total reward: 114.18
Episode: 3      Length: 277     Total reward: 356.82
Episode: 4      Length: 422     Total reward: 563.76
Episode: 5      Length: 288     Total reward: 430.46
Episode: 6      Length: 151     Total reward: 148.37
Episode: 7      Length: 332     Total reward: 430.34
Episode: 8      Length: 295     Total reward: 387.07
Episode: 9      Length: 267     Total reward: 366.18
Episode: 10     Length: 302     Total reward: 350.20
average score: 373.7496449160432
================================================
```

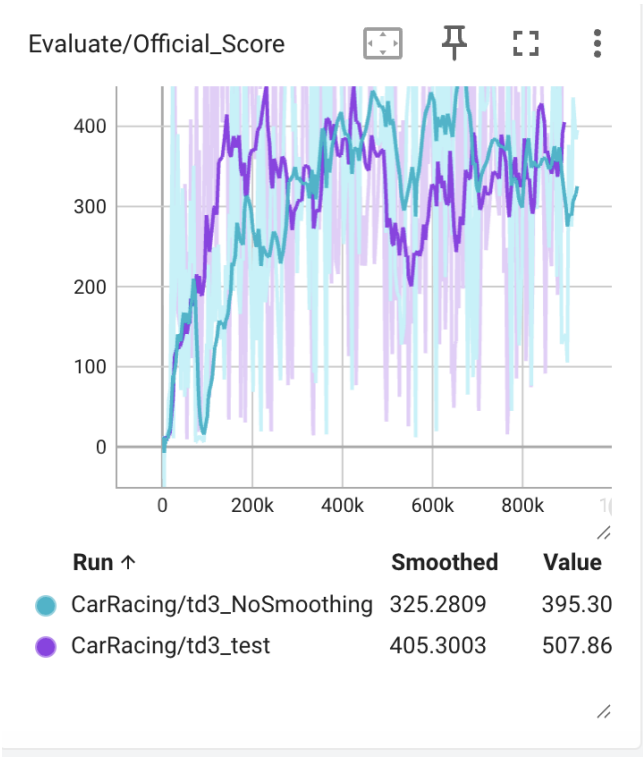




單一Q-value network存在嚴重的overestimate問題，問題，如果某 (state, action) 偶然給出了一個偏高的估值，這個錯誤會在訓練中被bootstrap並不斷放大，導致導致Actor學到很差的策略。 而TD3訓練兩個獨立的Q-value network。在計算 q_target 時，它會取兩者中的最小值。因為兩個network不太可能在同一個地方同時高估，取最小值能抑制overestimate的問題。 照理論來說，照理論來說，TD3（紫）應該會比Single Q（黑）的曲線穩定，並收斂速度更快。 從我的實驗成果來看，可能train的step較少有點看不出來，但總體來說TD3比Single Q還好一點。

2. **Screenshot of Tensorboard training curve and compare the impact of enabling and disabling target policy smoothing in TD3, and explain.**

```
Episode: 1      Length: 999     Total reward: 842.08
Episode: 2      Length: 999     Total reward: 879.17
Episode: 3      Length: 999     Total reward: 858.76
Episode: 4      Length: 703     Total reward: 929.60
Episode: 5      Length: 999     Total reward: 849.66
Episode: 6      Length: 999     Total reward: 840.77
Episode: 7      Length: 356     Total reward: 431.93
Episode: 8      Length: 999     Total reward: 883.97
Episode: 9      Length: 387     Total reward: 481.35
Episode: 10     Length: 999     Total reward: 864.66
average score: 786.1956564757758
----------------------------------------------------
```



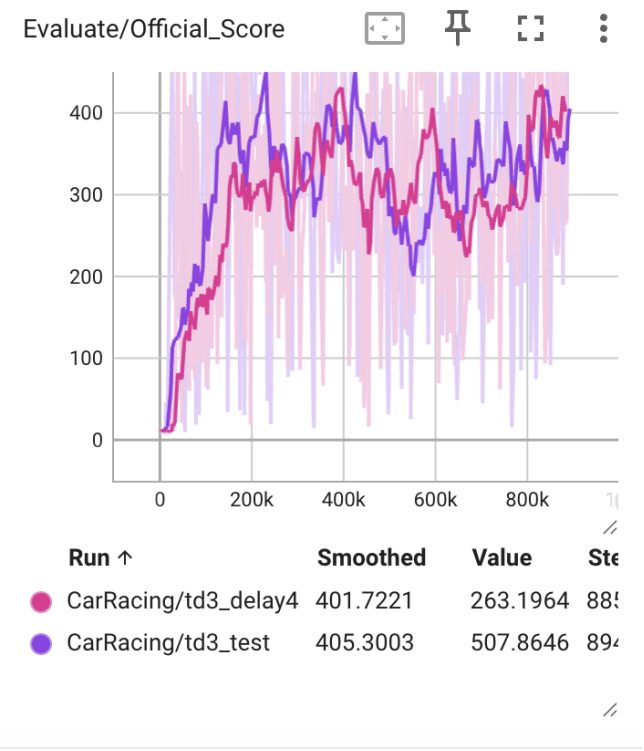
在 Actor-Critic 學習中，Critic(Q-value network)很容易Overfitting。它可能會在Q value上學習到一些非常狹窄的「假高峰」。Actor（Policy network）會很快地學會去利用這個「假高峰」，導致策略變得非常不穩定。

Policy Smoothing 會在計算 q_target 時，故意在target Actor產生的a_next上加入少量的噪音（policy_noise）。這會使a_next動作的「周圍」進行採樣，強迫Critic學習一個更平滑的Q-value。並使Actor的學習目標（Q 函數）更加穩定和可靠。

照理來說，原本有smooth的曲線（紫）應該要比沒smooth（藍）的曲線更穩定，且分數更高。但可能我訓練步數不夠，效果不夠明顯，並且暫時沒smooth的分數來略高。

3. **Screenshot of Tensorboard training curve and compare the impact of delayed update steps and compare the results, and explain.**

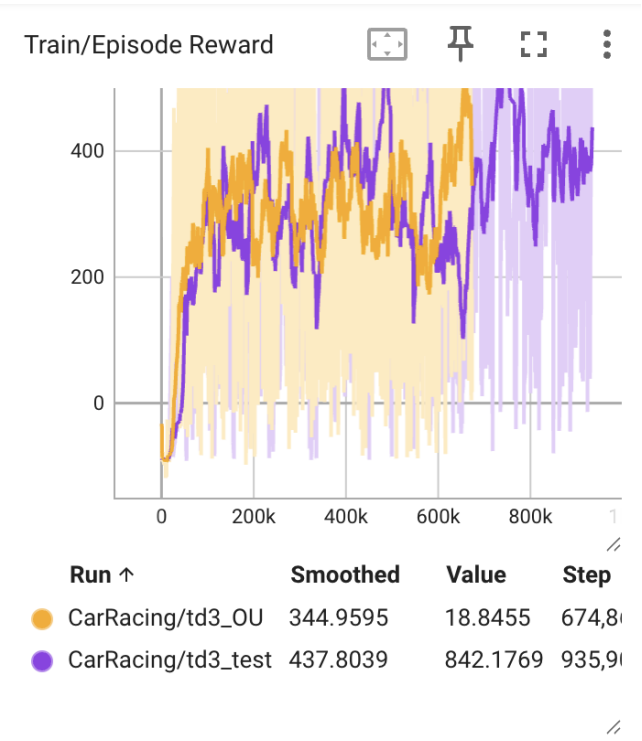
```
Episode: 1      Length: 999     Total reward: 808.45
Episode: 2      Length: 999     Total reward: 830.15
Episode: 3      Length: 999     Total reward: 838.22
Episode: 4      Length: 556     Total reward: 689.27
Episode: 5      Length: 174     Total reward: 175.48
Episode: 6      Length: 999     Total reward: 832.04
Episode: 7      Length: 644     Total reward: 617.32
Episode: 8      Length: 999     Total reward: 806.15
Episode: 9      Length: 344     Total reward: 320.15
Episode: 10     Length: 999     Total reward: 844.83
average score: 676.2055233447203
================================================
```

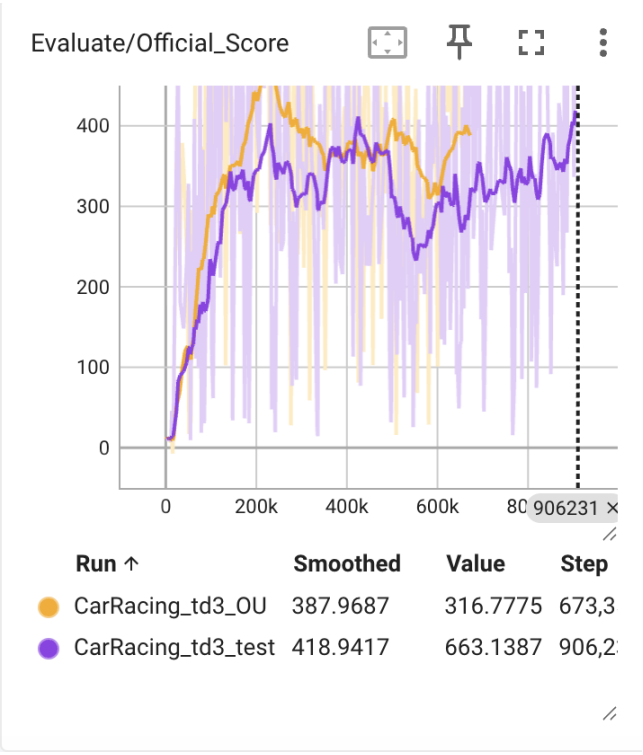


TD3 delay讓Actor(Policy Network)的更新頻率低於Critic(Q-value network)的更新頻率。透過設定 `update_freq: 2`，我們等於是讓Critic先更新2次Q-value，變得更準確穩定之後，才讓Actor的Q-value更新一次Policy。因此照理來說delay多的的曲線會更加穩定些，從我的實驗結果來看雖然沒到很明顯，但delay 4（紅）的還是有比delay 2（紫）的震盪小一點。

4. **Screenshot of Tensorboard training curve and compare the effects of adding different levels of action noise (exploration noise) in TD3, and explain .**

```
Evaluating...
Episode: 1      Length: 537     Total reward: 598.51
Episode: 2      Length: 487     Total reward: 660.10
Episode: 3      Length: 404     Total reward: 392.93
Episode: 4      Length: 609     Total reward: 718.41
Episode: 5      Length: 287     Total reward: 308.86
Episode: 6      Length: 358     Total reward: 413.14
Episode: 7      Length:  94     Total reward: 29.93
Episode: 8      Length: 477     Total reward: 578.01
Episode: 9      Length: 811     Total reward: 918.80
Episode: 10     Length: 911     Total reward: 864.07
average score: 548.2762792697417
```





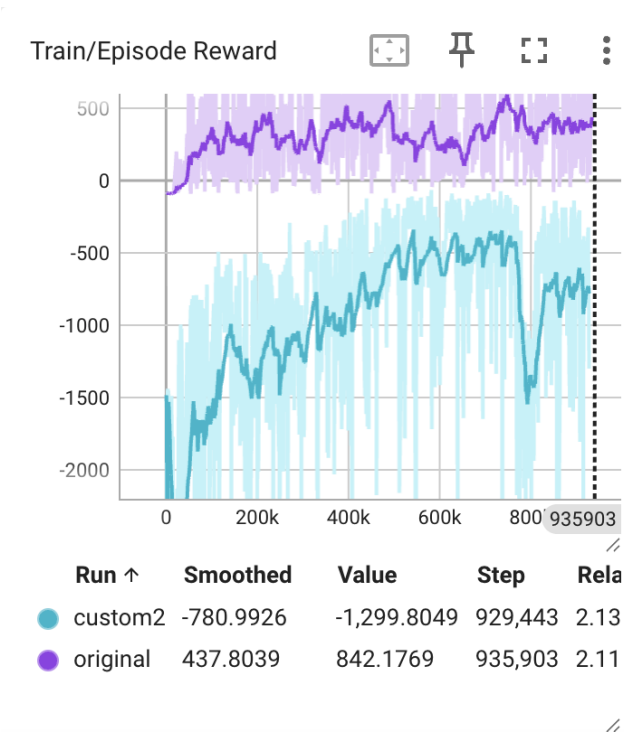
Gaussian Noise產生的噪音在每一步都是完全獨立的，前後完全不相關。

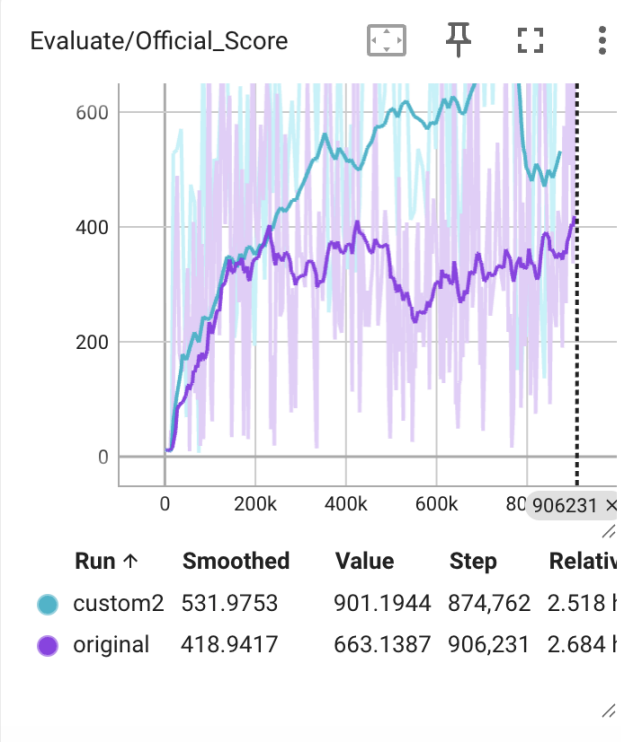
OU Noise是與時間相關的噪音，下一步的值是上步加上擾動，產生的噪音是平滑的，有助於像這個有慣性環境的遊戲像這個有慣性環境的遊戲更有效率地去探索。

以理論來說，OU Noise因為其平滑和時間相關的連續特性，更有可能產生像是「連續向左轉5步」 這種有意義的探索動作，更容易讓 Agent發現像是成功轉彎的動作，比Gaussian Noise這種完全隨機的動作更快脫離像直線撞牆時的狀況。 結果上OU noise（藍）較Gaussian Noise（紫）好一點。

5. **Screenshot of Tensorboard training curve and compare your reward function with the original one and explain why your reward function works better ?**

```
Episode: 1      Length: 931     Total reward: 906.80
Episode: 2      Length: 832     Total reward: 916.70
Episode: 3      Length: 827     Total reward: 917.20
Episode: 4      Length: 991     Total reward: 900.80
Episode: 5      Length: 967     Total reward: 903.20
Episode: 6      Length: 975     Total reward: 902.40
Episode: 7      Length: 750     Total reward: 924.90
Episode: 8      Length: 847     Total reward: 915.20
Episode: 9      Length: 845     Total reward: 915.40
Episode: 10     Length: 885     Total reward: 911.40
average score: 911.3999999999847
================================================
```





我的custom公式：`reward -= 0.1 * grass_pixel_count` 很單純地就是懲罰車子開到草地上，開到草地上時會減掉0.1*草地格子的分數，這會鼓勵車子開到路上，而從evaluate reward來看，我的reward funciton（藍）明顯很快就有效果，超越原本的td3_test(紫)。