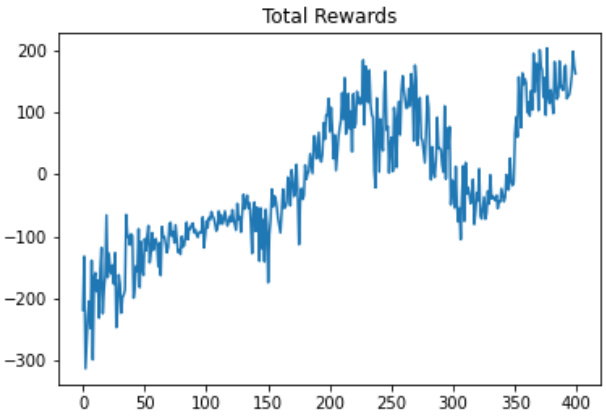
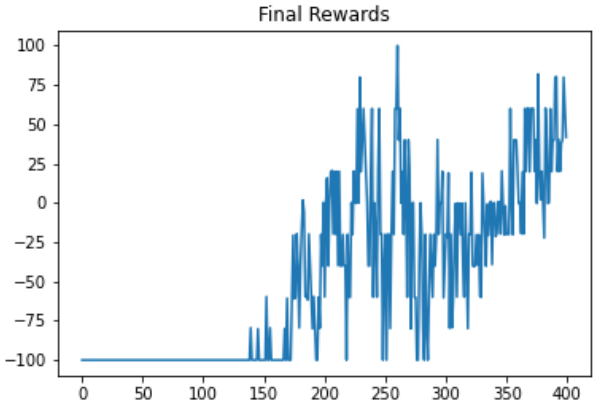
學號：b07901014　　系級：電機二　　姓名：陳希格

1. (20%) Policy Gradient 方法

   a. 請閱讀及跑過範例程式，並試著改進 reward 計算的方式。

   b. 請說明你如何改進 reward 的算法，而不同的算法又如何影響訓練結果？





改進 reward:

(1) 用老師的 Tip 2: Assign Suitable Credit(discount reward) with γ = 0.99

(2) 同時把 reward 的正規標準化拿掉，換成減掉 baseline(baseline = 前面幾個 epoch 的 np.mean(rewards) 的加權平均)

(3) 把改 optimzer；從 SGD 改成 adam(lr=0.01)

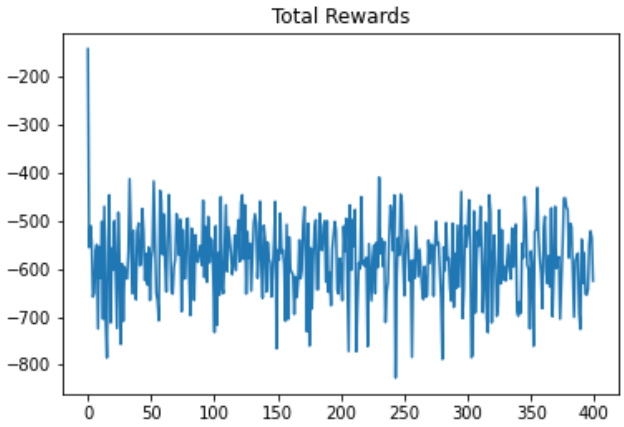
改進的算法比原本的好很多(可以從上圖看到)，尤其是(1)影響很大。

。

2. (30%) 試著修改與比較至少三項超參數（神經網路大小、一個 batch 中的回合數等），並說明你觀察到什麼。

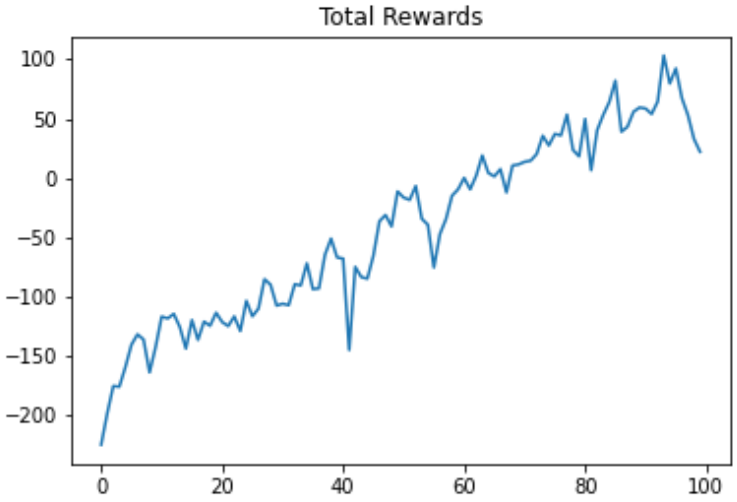
以下改動都是基於第一題改進後的方法

(1) optimzer；adam 換成 SGD



可以看到 SGD 完全 train 不起來

(2) 一個 batch 中的回合數 : 從 5 換成 20(epoch=100)

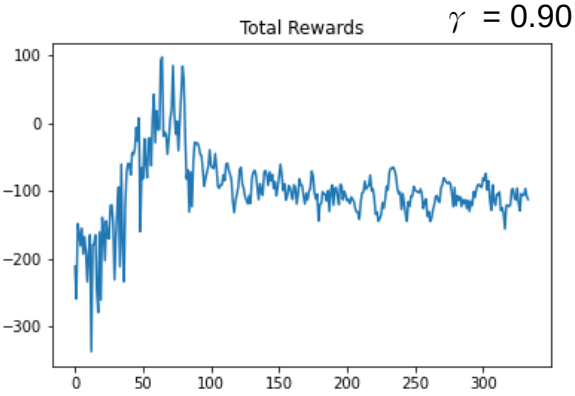


可以發現一個 batch 中的回合數變大可以明顯增加穩定性，reward 比較不會太大的變動。

(3) discount reward 中的:從 0.99 改成 1 和 0.90

γ = 1

γ = 0.90



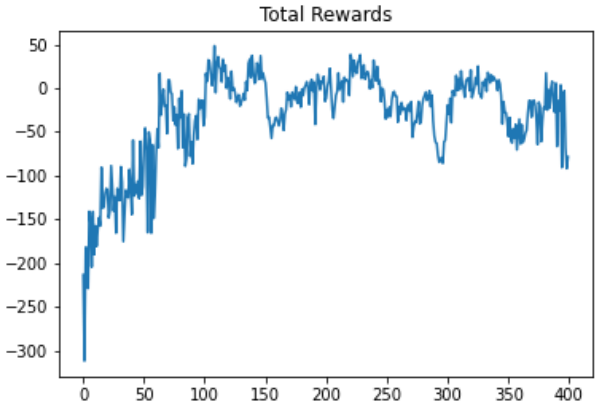
看起來 γ 過小或是過大都是不行 train 起來的，但詳細原因還不是很清楚。

3. (20%) Actor-Critic 方法

   a. 請同學們從 REINFORCE with baseline、Q Actor-Critic、A2C 等眾多方法中擇一實作。

   b. 請說明**你的實做與前者（Po**licy Gradient）的差異。

實作: REINFORCE with baseline(投影片上的方式)



從結果來看，REINFORCE with baseline，train 不起來，因為他沒有學到正確的降落(final 分數都是 0 左右)會有 100 分這件事。這可能需要 expert 來示範給他看，也就是 invserse reinforcement learning 才能解決。

4. (30%) 具體比較（數據、作圖）以上幾種方法有何差異，也請說明其各自的優缺點為何。

   從結果來看增加一個 batch 中的回合數是可以增加穩定度，但所需時間較久。而 lr 和 optimzer 都會影響穩定度。

   從第一題的圖中可以看到一個有趣的現象，reward 到 100 左右一段時間後又會調到-100，之後再慢慢 train 上去，這個現象在第二題的第三小題也可以觀察到。推測原因是有沒有成功降落分數影響很大，而機器無法學會如何好好的降落。如第三題所說，reward 的 function 可能需要 invserse reinforcement learning 才能解決。