CEDL-HW2 Policy Gradient

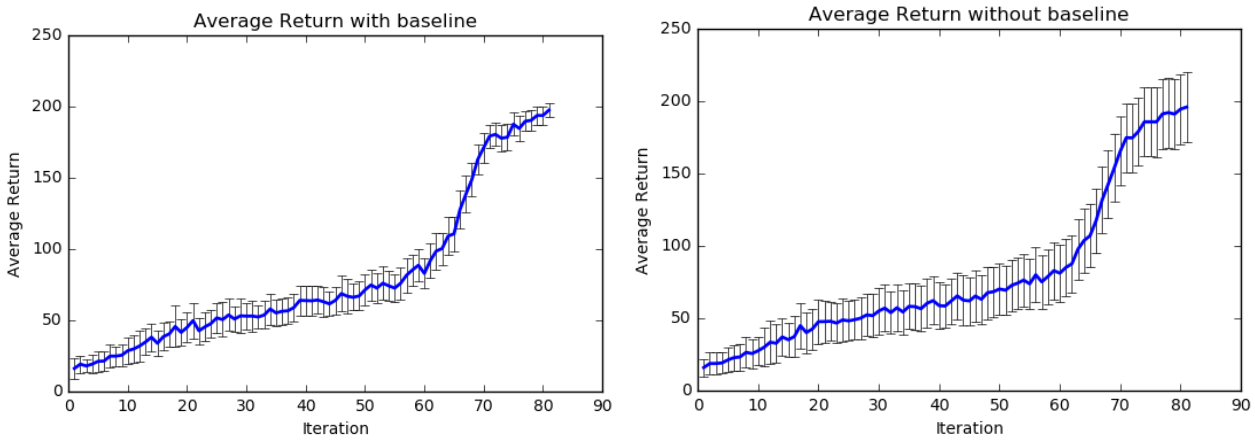
姓名：李冠毅　學號：104064510
姓名：李季紘　學號：(交大)0556083

簡介

本次作業目標為針對 OpenAI gym 的 CartPole 以 Policy Gradient 方式進行 Reinforcement Learning，在 Training 的過程中，以 Markov Decision Process 方式將 Observation(或是 State)、Action 放入 2-layer neural network 中，求出 Surrogate loss 以及 Reward，雖後以 Tensorflow 的 AdamOptimizer 執行迴圈，直到對應的 Rewards 值超過定義的 195 後停止 Training。

由於 Problem 1~4 是 Coding 部分的內容，因此本篇報告僅對於 Problem 5、6 以及其他的小問題作探討。

Problem 5：



圖一、Average Return 比較圖

上兩圖中的藍色線為 Average Return 值，而黑色線則表示 Stardard Deviation 範圍，左圖是有在 Policy Gradient 中加入 Baseline，而右圖則無，可以看到 Standar Deviation 值具有明顯差異，換算則 Variance 的話，有加入 Baseline 大約可以減少 Variance 約300～400左右，而在 Variance 值減少的情況下，原本可預期減少 iteration 數，在實驗中也曾測到 Iteration 數減少約10～20，但是因為每次執行的結果都不同，所以這裡給的數值只是大概值。

Problem 6：

針對 Advantage 進行 Normalization 的話能夠穩定 Rewards 中的 variance 大小的影響，進一步讓 Iteration 數趨於穩定，原預期此一步驟可讓 Gradient 趨於穩定，然而經多次實驗後卻發現 Iteration 數量不減反增，加入 Normalization 僅能讓加入 Baseline 的因素影響減少而已，因此判定 Normalization 可讓 Training 過程穩定。

Problem 2：

在計算 Surrogate loss 部分，原預期是按照公式計算 Likelihood Ratio，不過由於 Tensorflow 的 AdamOptimizer 是以 Minimize loss 的方式計算 Gradient，因此需要在公式加上負號才能運作，事實上程式也是在 Surrogate loss 加上負號後，才有可能 Converge。