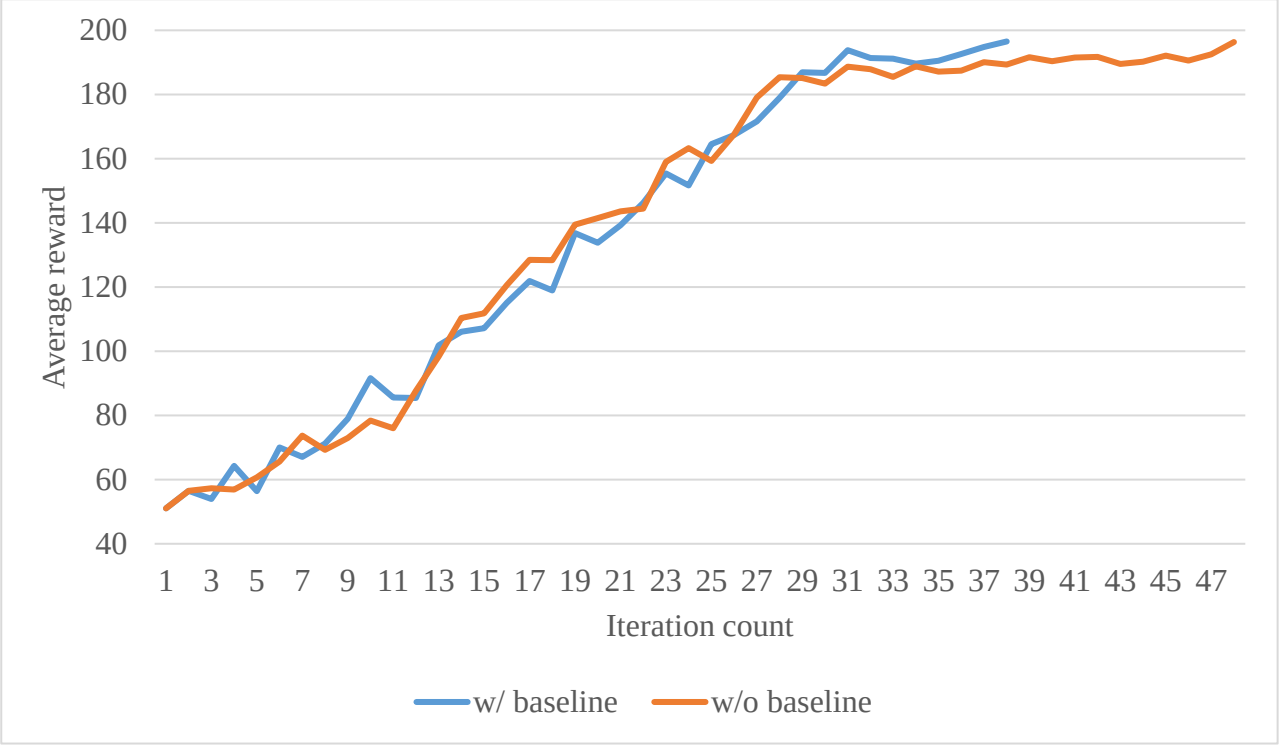
TheCEDL Homework 2

李青峰、陳金博、羅羿牧

1. 比較加入 baseline 前後的效能比較

加入 baseline 的目的是在不改變 policy gradient 的期望值的前提下降低變異量（也可以說是 noise），減少訓練所需時間。

圖（一）顯示加入 baseline 前後兩者 average return (AR) 的曲線變化，在第 29 次 iteration 前兩者的 AR 持續上升，起幅互有高低，在第 29 次之後兩者上升幅度趨緩，但加入 baseline 的 model 收斂需要的次數較少，未加入 baseline 的 model 則處於瓶頸，需要較多的 iteration 才能收斂，顯示加入 baseline 有助於加快收斂並提升效能。

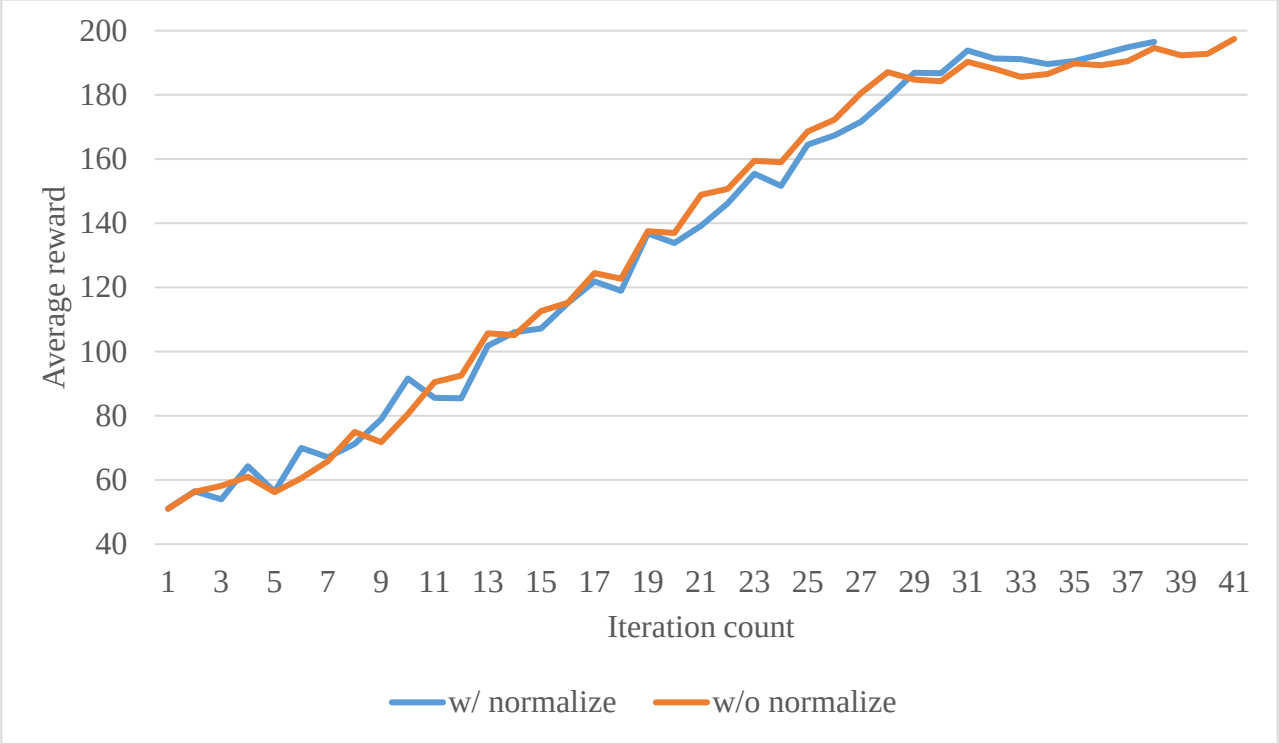


圖（一）

2. 解釋需要對 advantage 進行 normalization 的原因

我們發現這步驟有點類似 Batch Normalization 的技巧，只不過在這裡是針對 advantage function 做的。多了這步驟使得 policy gradient vector 中值的平均值固定為 0、variance 為 1，與 vector 的長度無關。這避免了 surrogate loss 受到模擬的時間長度增加影響而導致期望值驟降（因為此時 policy gradient vector 的平均值會提高），同時導致模擬後期的不同 actions 對 surrogate loss 的影響微乎其微。這樣會使得學習目標除了延長成功模擬的時間之外，過分專注在前期的表現，而不太去「懲罰」模擬後期導致失敗的 actions（然而直覺上「修正錯誤的行動」才是訓練過程中更重要的目標），因此讓訓練的效果打折扣，花的時間也更多。

圖（二）顯示 normalization 加快了收斂的速度的結果。



圖（二）

3. 貢獻

李青峰：安裝環境、解題、Problem 6

陳金博：實驗數據、解題、Problem 6

羅羿牧：Report 撰寫

註:為了方便實驗比較,我們在 HW2_Policy_Graident.ipynb 檔案中 In[2] 加入一行 env.seed(0)，將 OpenAI gym 的亂數種子固定。刪除這行對實驗結果可能有些許影響，但不影響結論。