# 111學年度輔仁大學資訊工程學系畢業專題

# 利用演化多工增強式學習解決模擬控制機器人問題

# Evolutionary Multitask Reinforcement Learning for Simulated Robot Control Problem

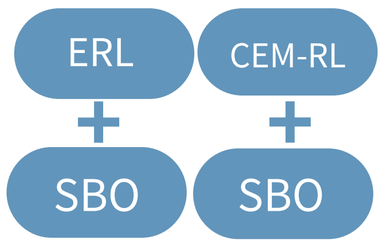
**組別**: A06　**指導教授**: 廖容佐 博士　**組員**: 資工四甲 羅鈺婷 / 資工四甲 葉承翰 / 資工四甲 楊謹芳 / 資工四乙 劉懷萱

## 目標/ 動機

我們期望使用Machine Learning來實現仿生機器人的控制，而因為Reinforcement Learning已經在控制問題的領域有顯著的成果，因此我們以Reinforcement Learning作為實驗方向。Reinforcement Learning是一種目標導向的學習方法，目的在於，透過與環境互動的過程中，獲得的各種獎勵或懲罰，學會如何做決策，使決策能達到最佳的學習成效。

## 專題概念

雖然Reinforcement Learning的方法已經在各種控制問題有很好的成果，但卻存在數據稀疏問題而多任務學習可以透過來自其它相關學習任務來幫助解決強化學習的數據稀疏問題。我們使用的實驗方法，是基於論文Evolution-Guided Policy Gradient in Reinforcement Learning和CEM-RL: Combining evolutionary and gradient-based methods for policy search中 強化學習結合演化學習:也就是ERL演算法和CEM-RL演算法為基礎。在了解原論文的做法後，我們嘗試加入演化多工 : symbiosis in biocoenosis optimization (SBO)演算法，主要是試著讓兩個任務同時進行訓練，並且透過交換offspring中的個體進行多任務間的學習，期望能夠提升效能讓訓練達到更好的學習效果。



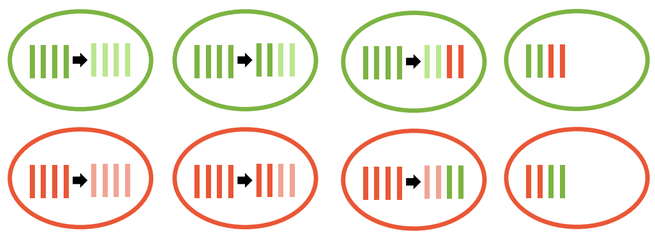

### ● 實驗平台

openAI gym

MuJoCo
Advanced physics simulation

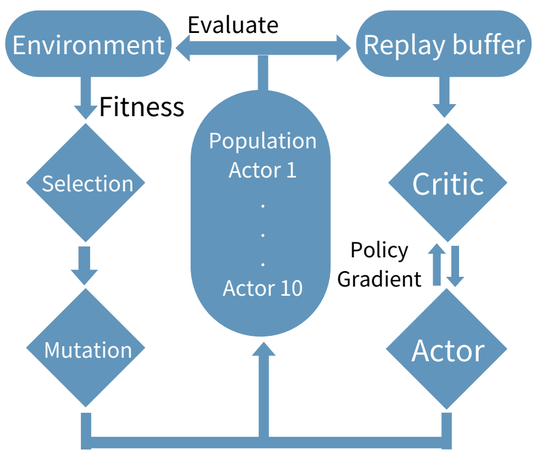
### ● SBO

SBO 的運作如下圖所示。首先會有許多問題，每個問題都有 EA 去解決，在 EA 解決的過程會有 Population，Population 中的每一個個體都代表著一種解，然後這些解會繁衍出子代。繁衍出子代之後，就開始進行資訊的傳遞。在 SBO 中，所謂的資訊傳遞就是透過取代子代來達成。至於子代取代的比例就要由 Transfer Rate 來決定。子代交換完之後，就會進行生存競爭，目的是將整體 Population 推向好的解。然後就照著繁衍、資訊傳遞、生存競爭這樣的流程持續進行下去。



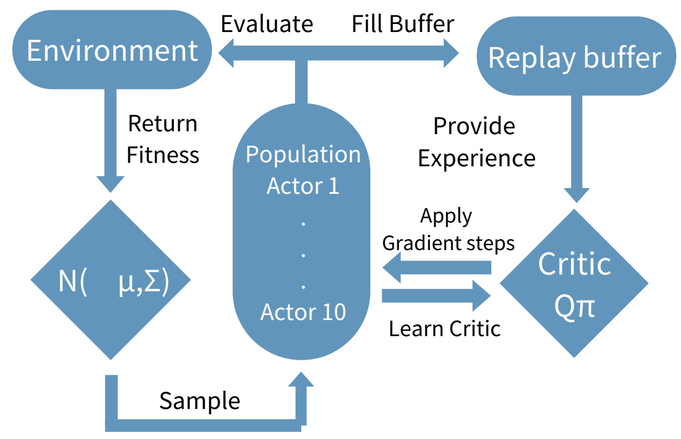
### ● ERL

在ERL中，在每次generation會計算population中actor的fitness，然後selection operator會根據fitness從population中選出parent進行crossover and mutation，成為offspring，而原本前二高fitness之actor則會保留為精英，並不受crossover and mutation步驟的影響。基於上述這種方法用來生成各種經驗來訓練 RL agent，並定期將 RL agent轉移到 population中以將梯度資料注入




### ● CEM-RL

CEM-RL 大致可分為3個part。 cem、ddpg/td3跟evaluation。每次generation會把population分為半。其中一半會先計算fitness而另一半會先被拿來更新critic再去計算fitness，我們會取fitness較高的一半當作parent透過高斯分布再去產生新的population。




## 成果

| | ERL(SAC) | | | ERL(SAC)-SBO | | |
|---|---|---|---|---|---|---|
| | Mean | CV. | Median | Mean | CV. | Median |
| HalfCheetah-v2 | 6417.00 | 13.60% | 6399.84 | 6831.645 | 10.38% | 7029.633 |
| Swimmer-v2 | 240.48 | 50.99% | 285.97 | 269.9592 | 39.04% | 276.035 |
| | ERL(SAC) | | | ERL(SAC)-SBO | | |
| | Mean | CV. | Median | Mean | CV. | Median |
| Walker2d-v2 | 1891.28 | 70.20% | 1355.20 | 1014.22 | 7.73% | 1003.13 |
| Ant-v2 | 1071.90 | 10.78% | 1040.73 | 1148.01 | 10.33% | 1145.76 |

上面是 SAC(前半部) 和 SAC結合SBO後(後半部) 的結果，可以看到在HalfCheetah-v2 和 Swimmer-v2 的組合中，結合 SBO 後兩者的 reward 比較高。而 Walker2d-v2 和 Ant-v2 的組合再結合SBO後，只有 Ant-v2 的 reward增加，Walker2d-v2 的 reward 反而下降了一點。

## 參考資料

Alo¨ıs Pourchot and Olivier Sigaud. CEM-RL: Combining Evolutionary and Gradient-Based Methods for Policy Search .The sixth International Conference on Learning Representations, 2019

S. K. Kagan Tumer. Evolution-Guided Policy Gradient in Reinforcement Learning. 32nd Conference on Neural Information Processing Systems. 2018

R.-T. Liaw and C.-K. Ting. Evolutionary manytasking optimization based on symbiosis in biocoenosis. In Proceedings of The Thirty-Third AAAI Conference on Artificial Intelligence, 2019.