多臂強盜（Multi-Armed Bandit）演算法比較分析

Lo C

April 21, 2025

簡介

多臂強盜問題 (Multi-Armed Bandit, MAB) 是強化學習中平衡「探索 (Exploration)」與「利用 (Exploitation)」的典型問題。本報告探討四種常見 MAB 策略：Epsilon-Greedy、UCB、Softmax 與 Thompson Sampling，並以程式模擬與圖表說明各自效能與差異。

1 演算法公式與邏輯

1.1 Epsilon-Greedy

$$\text{With probability } \varepsilon: \text{ choose a random arm}$$
$$\text{With probability } 1-\varepsilon: \arg\max_a Q_t(a)$$
$$Q_{t+1}(a) = Q_t(a) + \frac{1}{N_t(a)}(R_t - Q_t(a))$$

ChatGPT Prompt: 請解釋 Epsilon-Greedy 演算法中如何透過 ε 在探索與利用之間取得平衡，並產生對應的 Q 值更新公式。

1.2 UCB (Upper Confidence Bound)

$$A_t = \arg\max_a \left[Q_t(a) + \sqrt{\frac{2\ln t}{N_t(a)}} \right]$$
$$Q_{t+1}(a) = Q_t(a) + \frac{1}{N_t(a)}(R_t - Q_t(a))$$

ChatGPT Prompt: 請生成 Upper Confidence Bound（UCB）演算法的選擇策略公式，並解釋為什麼置信上界能夠提升探索效率。

1.3 Softmax

$$P(a) = \frac{e^{Q_t(a)/\tau}}{\sum_b e^{Q_t(b)/\tau}}$$
$$Q_{t+1}(a) = Q_t(a) + \frac{1}{N_t(a)}(R_t - Q_t(a))$$

ChatGPT Prompt: 請說明 Softmax 策略如何使用溫度參數 τ 來調整選擇動作的機率分佈，並產生機率公式。

1.4 Thompson Sampling

$$\theta_a \sim \text{Beta}(\alpha_a, \beta_a)$$
$$A_t = \arg\max_a \theta_a$$
$$\alpha_a = \alpha_a + R_t, \quad \beta_a = \beta_a + (1 - R_t)$$

ChatGPT Prompt: 請說明如何使用 Beta 分布作為先驗來更新每個 arm 的成功率估計，以實作 Thompson Sampling。

2 程式碼概述與模擬圖表

程式語言與設定 本模擬使用 Python 編寫，重複試驗輪數 $T = 1000$，臂數 $k = 4$，並為每種策略記錄其累積回報。

Python 繪圖程式片段 以下為統一執行四種演算法並產生圖表之簡要：

```
results = {
    "Epsilon-Greedy": epsilon_greedy(),
    "UCB": ucb(),
    "Softmax": softmax(),
    "Thompson␣Sampling": thompson_sampling()
}
for label, reward in results.items():
    plt.plot(reward, label=label)
plt.title("MAB策略累積回報比較")
plt.xlabel("輪數")
plt.ylabel("累積回報")
plt.legend()
plt.savefig("mab_plot.png")
```

Figure 1: 四種多臂強盜策略在 1000 輪試驗下的累積回報比較

實驗圖表

3 結果分析與複雜度比較

3.1 效能總結

- **Thompson Sampling**：表現最穩定，早期可辨識最優臂並快速收斂。
- **UCB**：中期策略效率高，逐漸收斂，具理論保證。
- **Softmax**：需依據溫度參數微調，對初始估計敏感。
- **Epsilon-Greedy**：簡單直覺但長期不易最適，固定 ε 無法適應動態環境。

3.2 空間與時間複雜度比較

策略	時間複雜度	空間複雜度	參數依賴性	收斂速度
Epsilon-Greedy	$\mathcal{O}(1)$	$\mathcal{O}(k)$	ε 需設計	中
UCB	$\mathcal{O}(1)$	$\mathcal{O}(k)$	無參數	快
Softmax	$\mathcal{O}(k)$	$\mathcal{O}(k)$	溫度 τ	中
Thompson Sampling	$\mathcal{O}(k)$	$\mathcal{O}(k)$	無需參數	快

Table 1: 各演算法複雜度與參數依賴比較

4 應用與策略選擇建議

- **Epsilon-Greedy**：適合教學與穩定環境初學者。
- **UCB**：適合已知 horizon 長度與理論要求的場景。
- **Softmax**：可套用於動態策略切換場景，如推薦系統。
- **Thompson Sampling**：推薦用於真實應用、即時決策與回報分佈不穩定場景。

結論

本報告以公式、提示語、程式與圖表全面比較四種 MAB 策略，結果顯示 Thompson Sampling 為整體最穩定的策略。根據使用場景與可接受的資源配置，其他演算法亦各有應用價值。