基於階層式強化學習的電網拓撲控制最佳化研究

Optimization of Power Grid Topology Control Based on Hierarchical Reinforcement Learning

一、摘要

電力網路在現代社會中扮演關鍵基礎建設的角色，隨著可再生能源使用率升高、電力使用量不斷成長，電網的即時控制與安全營運越顯複雜。強化學習（Reinforcement Learning, RL）因其能夠在高維度、動態環境中學習對應之最佳控制策略，近年來開始在電網控制領域受到關注。然而，電網拓撲控制（Topology Control）屬於高維度離散動作空間，再加上系統不確定性高，導致傳統的平坦式 RL 很難直接有效訓練。本研究計畫擬結合階層式強化學習（Hierarchical RL, HRL）之概念，並配合交替（區塊）座標梯度更新（Block Coordinate Gradient Update）方法，嘗試在電網拓撲控制問題中找出更高效且可解釋之控制策略。透過程式實作並進行模擬實驗，將與基準（如貪婪式搜尋或單層式 RL）方法進行比較，期望能以更低的訓練成本及更高的可擴充性，達到有效率的電網安全運轉。

二、研究動機與研究問題

（一）研究動機

- **電網需求與挑戰**

在電力系統的即時調度中，系統營運者必須確保整體供需平衡並避免輸電線路過載。隨著可再生能源（如風能與太陽能）的滲透率日益增加，電力生產及負載變動具有更大的不確定性；同時，輸電網路設備老化、跨區域互聯等因素也提升了傳統規劃工具的複雜度。

- **高維度動作空間**

  配電/輸電拓撲控制不僅面臨輸電線數量多，且每個變電站（Substation）內部也可能有多條線路、母線切換組合，如此複雜的組合行為往往導致指令空間呈現指數爆炸（Combinatorial Explosion）。

- **階層式強化學習的潛力**

  我們認為可以透過分層策略處理不同層級（例如：中層選擇變電站、底層決定該變電站母線配置），可使得動作空間有效被切塊 (Block) 並進行交替更新，提升學習效率並兼具可解釋性。

（二）研究問題

1. 如何利用階層式強化學習（HRL）有效分解電網拓撲控制的高維度動作空間？

2. 在中層與底層之間，是否能以 Block Coordinate Gradient Update 的概念，交替訓練不同階層的策略（Policy），並保證最終收斂至合理的解？

3. 與傳統的單層式 RL 或貪婪式搜尋相比，階層式做法能否在因應隨機故

障、能源變動等真實場景下展現更好的效率與穩定性？

三、文獻回顧與探討

（一）電網拓撲控制與 RL 相關研究

- **電網控制的挑戰**

根據文獻 Marot[1]等人指出，電網控制面臨的挑戰主要來自於現代能源結構和技術要求的快速變化。隨著可再生能源的普及，供電的不確定性大幅增加，例如風力和太陽能的發電量因天氣條件而異，這導致了電力生產與消費之間的波動性加劇。傳統的電網控制方法已逐漸無法滿足這種複雜的需求，需要引入更智慧化的解決方案。

此外，電網的運行需要應對突發事件，例如線路超載可能引發的連鎖停電事件，這要求在極短的時間內進行有效的決策和調整，人類操作員很難在有限的時間內處理這些問題。因此，自動化和智慧化的電網運營系統逐漸被重視，尤其人工智慧與強化學習（RL）技術提供了嶄新的解決思路。

- **Learning to Run a Power Network (L2RPN)競賽**

為了促進 RL、AI 與電網相關的發展，L2RPN[2]是一種國際型競賽，旨在探索人工智慧（AI）和強化學習（RL）在電網控制中的應用潛

力。競賽的核心是開源程式碼的 Grid2Op 框架，這是一個模擬電網運行的工具，參賽者可在其中測試不同的策略。

比賽設計了兩個主要評比方式：適應性（Adaptability）和穩健性（Robustness）。適應性要求代理(agent)能夠應對不同的能源組合，特別是隨著可再生能源比例增加的情況。穩健性賽道則模擬了線路斷電等對手行為，挑戰代理在惡劣條件下維持電網穩定。但現存成果多針對縮減後的動作集合，或需要搭配啟發式規則，才有望在複雜網路上求得可行解。

（二）階層式強化學習（Hierarchical Reinforcement Learning, HRL）相關研究

- **HRL 基本概念**

HRL 是一種強化學習的擴展方法，其核心思想是將任務分解為不同層次的子任務，並為每個子任務設計單獨的學習和決策模組。這種方法模仿人類的決策過程，透過高層策略控制低層策略，便能更有效地處理複雜且多步驟的問題。

高層策略負責制定長期目標，例如「到達某個地點」; 低層策略則執行具體行動來實現高層目標，例如「向前移動」或「避開障礙物」。這樣的結構化學習方式，在面對大規模動作空間和長期依賴性的挑戰時特

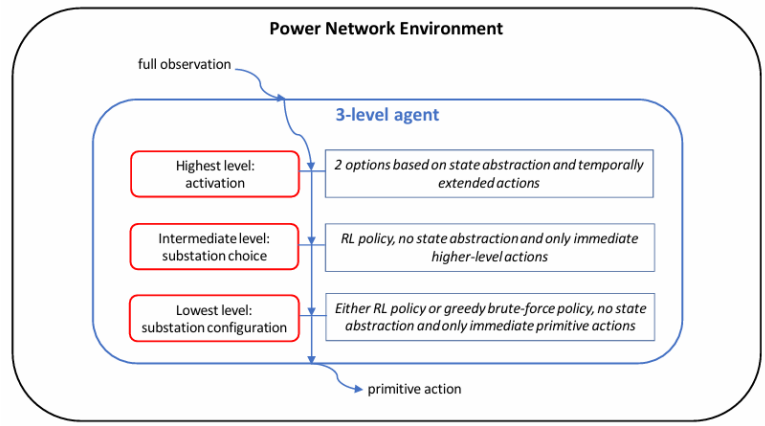
別有效率。

- **HRL 應用於電網相關研究**

根據 Manczak[3]等人的研究，HRL 已在電網控制領域展現顯著潛力。HRL 將複雜任務分解為多層次的子任務，能有效應對電網運營中的高維狀態和動作空間問題。

研究也提出了三層的架構(如圖一)，分別對應不同的決策層次：第一層是基於規則的策略，用於判斷當前是否需要採取行動（例如“保持現狀”或“提出拓撲更改”）；第二層採用強化學習（RL）算法，用於選擇需要調整的具體變電站；第三層則處理所選變電站的具體配置，確定最佳拓撲方案。Manczak[3]等人的研究中，第二層和第三層的策略可以分別使用不同的 RL 演算法進行訓練，例如，可以使用 PPO (Proximal Policy Optimization）或 SAC (Soft Actor Critic)等策略梯度方法來優化第二層選擇變電站的策略，並使用 PPO 或 SAC 來優化第三層的具體拓撲配置策略。這種設計允許針對不同層級的子任務選擇最適合的 RL 演算法，進一步提升了整體框架的效能。由於 PPO 和 SAC 在處理離散和連續動作空間方面都具有良好的表現，因此它們成為了 HRL 框架中常用的演算法選擇。 這樣的分層設計大幅減少了行動空間的規模，提升了模型在處理高維問題時的效率和穩定性。

實驗結果顯示，該三層 HRL 框架在不同的實驗場景中均表現出色。在較為簡化的環境中（無隨機線路中斷），HRL 代理的成功率接近100%，並且操作次數顯著減少，證明了該架構在低風險情境下的高效性。在包含隨機線路中斷的現實環境中，HRL 代理展現了良好的適應性和穩健性，成功在大多數測試場景中保持了電網的穩定運行，並明顯優於基於貪婪搜索或單層 RL 的代理。



(圖一) Manczak[3]等人所提出之 HRL 架構

（三）Block Coordinate Descent

Block Coordinate Descent 是一種廣泛應用於 Convex OptimizationNon-Convex Optimization 問題的數學方法，其核心思想是透過交替更新變數的分塊子集以逐步逼近全局或局部最優解。這種方法的分塊策略可以有效降低高維問題的計算複雜度，同時為解決結構化問題提供靈活性。

（四）PPO (Proximal Policy Optimization）及 SAC (Soft Actor Critic)

PPO 是一種由 Schulman[4]等人提出的高效強化學習策略梯度方法，

其具備穩健性與樣本效率，被廣泛應用於動態決策問題。PPO 透過引入「剪裁的代理目標函數」 (Clipped Surrogate Objective)，控制策略更新的幅度，防止模型因過大更新而崩潰，實現穩定的策略改進。相比於早期的信任域策略最佳化 (TRPO)，PPO 計算更簡單，實現更靈活，能適應具有高不確定性的動態場景。

SAC 是 Haarnoja[5]等人基於最大熵強化學習（Maximum Entropy Reinforcement Learning）的一種演員-評論家（Actor-Critic）方法，應用特別適用於連續動作空間，主要透過熵的正則化(使未來策略更具有探索性)、雙 Q 值網路(降低 overfitting)，以及離線學習(用舊數據進行訓練，比 PPO 更具樣本效率)

$$J(\pi) = \sum_{t=0}^{T} E\left[r\left(s_t, a_t\right) + \alpha H\left(\pi(\cdot \mid st)\right)\right]$$

根據 Tuomas Haarnoja[6] 等人提出的研究，在既有的無模型深度強化學習（RL）演算法所遇到的問題，SAC 可解決其所面對的兩項主要挑戰，分別為高的樣本複雜性，以及脆弱的收斂特性，相較其他的離散策略方法，SAC 在應對更加複雜以及龐大的資料時，都能表現相似性能。

在 Manczak[3]等人的階層式強化學習研究中，分別使用 PPO 與 SAC 兩種強化學習演算法來訓練中間層代理人，以評估其在高維離散動作空間中的表現。研究結果顯示，無論是穩定的環境或是更具挑戰性的

環境，PPO 皆比 SAC 更適合這類電網控制問題，因為 SAC 的學習過程不穩定且表現較差，而 PPO 因為其策略更新機制限制了策略變動的幅度 (clipped objective function)，提高了學習的穩定性，才能獲得較好的表現。基於 Manczak[3]等人的研究，本研究計畫在單層強化學習的訓練上，將參考其實驗結果，採用 PPO 作為主要學習演算法，以提高模型的學習穩定性與效能。

（五）小結

雖然 HRL 在電網拓撲控制中展現了良好的性能，但因電網拓撲控制涉及高維離散動作空間，上層策略的決策直接影響下層策略的訓練表現。若採分散式收斂的訓練模型，可能導致上下層之間的協調性不足，直觀上可能出現不一致的情況，從而影響整體性能並降低收斂效率。

我們想驗證透過把中層 (例如：選哪個變電站) 與底層 (例如：該變電站要切換至何種母線配置) 做區塊座標（Block）方式交替更新，類似在凸規劃中的 Block Coordinate Descent，設計一種在 HRL 框架中引入 BCD 的訓練流程，嘗試同時提升模型的訓練效率與全局收斂性能。

四、研究方法及步驟

本研究之流程將分為下列幾個階段，並在「階層式 RL 架構設計」與「演算法實作」中導入 **Block Coordinate** 更新機制之明確定義與演算法內容：

1. **問題及資料蒐集**

   - 使用公開的電網模擬套件（如 Grid2Op）以及 IEEE 14-bus、IEEEXX-bus 等不同規模的測試案例作為訓練與測試環境。
   - 針對各案例紀錄各條輸電線、變電站拓撲、限載容量與能源時序變化等資訊。

2. **階層式 RL 架構設計**

   - 最高層：採用活動門檻值（Activity Threshold）或類似 Option 的二元策略："目前是否需要進行拓撲操作？"。
   - 中間層 (中層)：選擇單一或多個變電站進行調整；輸入為完整觀測（或抽象化的 state）後，輸出一個 substation 編號。
   - 最低層 (底層)：針對中層輸出的變電站，決定母線配置或線路切換組合。

   模型訓練時，參考 Block Coordinate Gradient Update：先固定底層策略，更新中層；再固定中層策略，更新底層，反覆交替到收斂。

3. **演算法實作與模擬實驗**

   - **實作工具**：Python（搭配如 RLlib、PyTorch 等），並利用 Grid2Op 進行環境互動。

- **比較方法**：
  - 貪婪式搜尋（一次性在所有可行拓撲裡擇優）
  - 單層式 PPO 或 SAC（不做階層分解）
  - 階層式 RL：
    1. 「PPO Substation + 底層貪婪」: 中層使用 PPO，底層使用貪婪搜尋。
    2. 「PPO Substation + PPO 底層」: 完全雙層 PPO。
- **評估指標**：
  - 成功維持電網連續運轉的平均時間或測試情境成功率。
  - 平均或總累積獎勵（如負載率分佈狀況、超載違規次數）。
  - 計算量（訓練時間/推論時間）及動作次數、拓撲變動之複雜度（如修改過的母線數量）。

Algorithm 1 階層式 RL 之 Block Coordinate 更新流程

Require: Initialize θ_{mid} (parameters for mid-level policy $\pi^{(mid)}$)
Require: Initialize θ_{low} (parameters for low-level policy $\pi^{(low)}$)
Set iteration counter k = 0

while not converged do
    k = k + 1

Phase A: Fix bottom layer, update mid layer

1. **Fix** θ_{low}
2. Using the environment (Grid2Op) + $\pi^{(low)}$ for bottom decisions,
   gather episodes data for mid-level training
3. **Update** θ_{mid} via policy gradient (e.g., PPO)
   → minimize θ_{mid} (θ_{mid} | θ_{low} fixed)

Phase B: Fix mid layer, update bottom layer

4. **Fix** θ_{mid}
5. Again, gather episodes data, but now we let $\pi^{(mid)}$ choose substation
   while exploring bottom-level configurations
6. **Update** θ_{low} via policy gradient (e.g., PPO)
   → minimize L_{low} (θ_{low} | θ_{mid} fixed)

Optional: check if improvement in reward or performance is below a threshold

If below threshold for consecutive steps, we can terminate early

end while

Return final (θ_{mid}, θ_{low})

4. **數據分析與結果比較**

   - 分析與比較各種架構在因應不同故障（Line Outage）、能源隨機變動下的表現差異。
   - 檢視階層式 RL 在不同網路規模下的可擴充性、訓練效率與穩定

度是否優於單層式方法。

5. **演算法修正與實例分析**

   - 若結果不達預期，將針對 Block Coordinate 更新機制、Reward 參數調整或子任務設定再次修正。
   - 成果可進一步應用於**真實或更大型**的電力系統測試案例，並探討其可行性。

五、預期結果

1. **高維度動作空間下的高效率學習**

   - 透過階層式強化學習，預期能在不犧牲性能的前提下，顯著降低搜尋或訓練時間。

2. **更具備解釋性的策略**

   - 不同層級策略可明確解釋：中層選哪個站、底層如何切換線路或母線分配等。

3. **應對真實場景的穩定度**

   - 在故障情境或能源間歇性下，期望階層式方法能更具穩健性，避免過度依賴短視（greedy）動作而導致崩潰。

4. **電網營運應用潛力**

    - 如能更好地在模擬中維持線路負載安全範圍，未來可移植至實務之決策支援平台，甚至發展至更複雜（多電壓層級或大量節點）的應用。

六、需要指導教授指導內容

1. **演算法設計與程式實作建議**

    - 針對 Block Coordinate Gradient Update 與階層式強化學習框架的實作細節、收斂條件檢核等。

2. **對真實電網或大型測試網之應用**

    - 可能在資料處理、參數設定、實際拓撲規劃限制上需要指導老師協助把關。

3. **成果應用建議**

    - 如何將研究結果延伸至跨區域電網或結合其他控制（如發電機排程等），擴大應用層面。

七、參考文獻

1. Marot, A., Donnot, B., Dulac-Arnold, G., Kelly, A., O'Sullivan, A., Viebahn, J., ... & Romero, C. (2021, August). Learning to run a power network

challenge: a retrospective analysis. In *NeurIPS 2020 Competition and Demonstration Track* (pp. 112-132). PMLR. Blazej Manczak. (2023).

2. ChaLearn. (n.d.). L2RPN challenge. Retrieved January 20, 2025, from https://l2rpn.chalearn.org/https://l2rpn.chalearn.org/

3. Manczak, B., Viebahn, J., & van Hoof, H. (2023). Hierarchical Reinforcement Learning for Power Network Topology Control. *arXiv preprint arXiv:2311.02129*.

4. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

5. Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018, July). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In International conference on machine learning (pp. 1861-1870). PMLR.

6. Tuomas H., Aurick Z., Pieter A., Sergey L. (2018) Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor

7. Huang, J. (2022). Distributed Algorithms for Finding Linear Arrow-Debreu Market Equilibria (Master's thesis, National Yang Ming Chiao Tung University).

8. Chen, P. A., Lu, C. J., & Lu, Y. S. (2022). An Alternating Algorithm for Finding Linear Arrow-Debreu Market Equilibria. Theory of Computing Systems, 1-18.

9. Chen, P.-A., Lu, C.-J., Lin, C.-C., & Huang, J. (2024). Finding linear Arrow-Debreu market equilibria: Trading fast convergence of the generalized block coordinate gradient projection for decentralization of the optimistic gradient descent ascent.

10. Beck, A., & Tetruashvili, L. (2013). On the convergence of block coordinate descent type methods. *SIAM journal on Optimization*, *23*(4), 2037-2060.

11. Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed

optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine learning, 3(1), 1-122.

12. Yoon, D., Hong, S., Lee, B. J., & Kim, K. E. (2021, May). Winning the l2rpn challenge: Power grid management via semi-markov afterstate actor-critic. In *International Conference on Learning Representations*.

13. Kelly, A., O'Sullivan, A., de Mars, P., & Marot, A. (2020). Reinforcement learning for electricity network operation. *arXiv preprint arXiv:2003.07339*. Malte Lehna, Clara Holzhüter, Sven Tomforde, Christoph Scholz.

14. Lehna, M., Holzhüter, C., Tomforde, S., & Scholz, C. (2024). HUGO--Highlighting Unseen Grid Options: Combining Deep Reinforcement Learning with a Heuristic Target Topology Approach. *arXiv preprint arXiv:2405.00629*.