Application Status of Deep Reinforcement Learning in Optimal Power Flow (OPF) Problem in Renewable Energy Power System

Shuai Yuan

School of international exchange, Guangzhou Maritime University, Guangzhou, China
YuanShuaishy@outlook.com

Abstract. This paper makes an in-depth analysis of the challenges faced by the optimal power flow (OPF) of the power system due to the high proportion of renewable energy access, which makes the uncertainty increase and the real-time requirements improve. Since the traditional iterative optimization method only relies on mathematical modeling (such as the interior point method) and it is difficult to meet the real-time decision-making needs of large-scale power systems, deep reinforcement learning (DRL) provides a key idea for solving the optimal power flow (OPF) problem with its dynamic adaptation and online rapid decision-making advantages. This paper first gives an overview of the optimal power flow related algorithms of the power system (such as single-agent, multi-agent deep reinforcement learning algorithms, and safe deep reinforcement learning algorithms), and then aims to solve the four defects of the traditional DRL algorithm design. From the dimensions of algorithm design, multi-agent collaboration, physical constraint fusion, and data and knowledge enhancement, this paper reviews the application of deep reinforcement learning in solving the optimal power flow problem, analyzes the technical bottlenecks, and provides future research directions, so as to provide reference for future related research.

Keywords: Deep reinforcement learning; Optimal power flow; Power system.

1 Introduction

With the promotion of the "double carbon" goal, the penetration of renewable energy (RES) such as wind power and photovoltaic in the power system continues to increase. According to the data of the national energy administration, the new installed capacity of distributed photovoltaic in China in 2024 will be three times that of 2021, accounting for 40% to 60% of the new installed capacity of photovoltaic in the whole year. However, due to the intermittent and fluctuating output of a high proportion of renewable energy, the optimal power flow (OPF) problem of power system has changed from deterministic optimization to uncertain dynamic decision-making [1]. If a large-scale power system fails, the loss will be great. However, the traditional iterative optimization method based on Newton method and interior point method is difficult to adapt to the rapidly changing reality. The research shows that the traditional iterative

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_77

method takes more than 10 minutes to process the random OPF with 1000 scenarios, and depends on the accurate system model, which is difficult to adapt to the real-time scheduling requirements [2]. Because of its flexible algorithm design, strong dynamic adaptability and fast response speed, DRL is applied to solve the OPF problem of optimal power flow.

The optimal power flow problem (OPF) is the core of power system operation. Its goal is to minimize the generation cost and meet the physical constraints (power balance, voltage/line power upper and lower limits). However, OPF has non convex nonlinear characteristics and belongs to NP hard problem. The traditional method is difficult to take into account the real-time and feasibility [3]. Deep reinforcement learning (DRL) can embed the "state action reward" mapping of optimal power flow OPF into neural network by constructing Markov decision process (MDP), which can realize the separation of off-line training and on-line decision-making, thus significantly reducing the burden of real-time computing. At present, DRL is the mainstream technology to solve such problems. However, the traditional deep reinforcement learning model (such as SAC) has four technical pain points that are difficult to solve when it is applied in power system, such as insufficient security assurance, poor interpretability, difficult training and limited generalization ability. However, the traditional deep reinforcement learning model (such as SAC) has four technical pain points that are difficult to solve when it is applied in power system, such as insufficient security assurance, poor interpretability, difficult training and limited generalization ability. Therefore, in recent years, relevant researches from the aspects of multi-agent DRL, physical informed DRL (pinn), security constraint DRL (sdrl) and data knowledge double enhanced DRL want to further solve the "dimensional disaster", the difficulty of constraint satisfaction, the lack of security and other defects of large-scale power system. For example, Zhou et al. Proposed a hybrid framework of "physical constraint preprocessing+intelligent algorithm optimization+data-driven correction" to solve the optimal power flow of high proportion renewable energy power system [2]; Ahmed sayed et al. Proposed a physically driven DRL algorithm integrating the differentiable holomorphic embedded power flow model (d-helm) for solving the optimal power flow of power system [3]; À lex tudoras-miravet and others solve partial differential equations of power system based on physical information neural network (pinn). In the transient analysis of IEEE 39 bus system with wind power, the prediction error of key variables is less than 2%, and only 50 groups of data are needed to reach the accuracy of 500 groups of data in the pure data-driven model [4].

This paper systematically combs the application status of DRL in OPF of renewable energy power system. Firstly, aiming at solving the four defects, it summarizes the relevant solutions, and then provides the research direction to provide reference for subsequent research.

2 Core Defects of Traditional DRL and Targeted Solutions in Existing Papers

The core defects of traditional DRL and the targeted solutions in existing papers traditional DRL applications (such as sac, PPO, ddpg, etc.) face four major challenges in solving OPF problems: insufficient security assurance, poor interpretability, difficult training and limited generalization ability [3]. Next, this paper discusses the existing technical solutions based on these four challenges:

2.1 Insufficient Security Assurance

Because the traditional DRL uses the "trial and error learning" optimization strategy, it is easy to produce dangerous actions that violate the hard constraints of power system (line transmission limit, etc.), and it is difficult to be directly applied to OPF. To solve this problem, the existing research proposes three kinds of constraint guarantee schemes:

• physical informed constraint embedding. Tudoras miravet et al. Used the physical informed neural network (pinn) for OPF warm start, added the penalty term of power flow equation in the loss function, and embedded the voltage amplitude boundary in the output layer, which increased the success rate of OPF solution by 51.9%, avoiding constraint violation from the source [4].

• security constraint DRL algorithm, Wu et al. Proposed pd-ppo algorithm, modeled OPF as constraint MDP, and calculated Lagrange dominance function through double comment network. The constraint satisfaction rate of IEEE 118 node system reached 97.93%, 14.6 percentage points higher than that of traditional penalty PPO [5].

• action mask and physical correction. Chen et al. Used energy storage capacity correction mechanism for continuous actions and masking mask for discrete actions, and the constraint satisfaction rate increased to more than 98% [5]; Wu et al. Pdo-amg method generates action mask based on power flow Jacobian matrix, and the line violation of activesg2000 system is reduced to 3.75×10^{-4}[5].

The above scheme has changed from "passive punishment" to "active constraint", which has well solved the core problem of traditional DRL security trial and error. Among them, pd-ppo and pdo-amg achieve the balance between security and economy through mathematical modeling and knowledge guidance. In large-scale systems, the constraint satisfaction rate exceeds 97%, which is more than 14 percentage points higher than the traditional punishment method [5].

2.2 Poor Interpretability

Because the traditional DRL neural network decision-making process is difficult to explain, the dispatcher cannot trust its "black box output". The existing research improves the transparency of model decision-making through knowledge enhancement and interpretation assistant technology, which is convenient for the dispatcher to understand:

• knowledge enhancement modeling. The ke-mappo model embeds explicit knowledge such as power balance and line load rate and tacit knowledge such as key factors of weak branches into the loss function as regular terms. The decision-making

in Tianjin 725 node grid conforms to the operation rules, and the network loss is reduced by 10.53% [6].

• DRL interpretation aided technology, Dai et al. Proposed an interpretation aided integer variable reduction method, used Gaussian mixture model to cluster the unit output distribution, combined with the improved decision tree to extract the "must start unit" and "must stop unit" rules, and transformed the complex strategy into logic that the dispatcher can understand [7].

• prior knowledge guides zoning, gcnii model identifies weak branches of the power grid, delimits the agent control area by extracting key factors from mutual information, makes the prior logic of multi-agent decision-making clearer, and reduces the load shedding rate of extreme scenarios to 0.095%[6].

The technology of knowledge embedding and interpretation AIDS has changed DRL decision from "non traceable" to "interpretable and verifiable". Ke-mappo makes the decision-making process fit the power system operation rules through explicit knowledge regularization; DRL interprets auxiliary reduction rules to transform complex strategies into rules understandable by dispatchers through clustering and decision trees [7].

2.3 Great Difficulty in Training

Traditional DRL relies on a large number of high-quality interactive data, while power system data is scarce and "trial and error cost is high", which leads to slow convergence of model training and easy to fall into the misunderstanding of local optimization. Existing research reduces the difficulty of training through the fusion of data mining and physical prior：

• data mining generates training samples, and the wcgan-gp method inputs the day ahead forecast data to generate the daily endogenous load fluctuation and n-1, n-2 fault scenario sets, which increases the amount of mappo algorithm training data by one order of magnitude, accelerates the convergence speed by 40%, and reduces the trial and error cost by 90% [6].

• physical priors accelerate convergence, pinn models such as tudoras miravet integrate the physical constraints of power flow to reduce ineffective exploration, and the calculation efficiency of OPF warm start is increased by 18.7% [4]; Wu et al. Pdo-amg method guided the movement exploration through the trend knowledge, and the training convergence speed was increased by 30% [5].

• multi agent parallel training. Zhou and other h-mappo algorithms adopt a multi-threaded asynchronous training architecture. The calculation speed of rts-gmlc system is 37-85 times higher than that of Matpower, which alleviates the problem of low training efficiency of single agent [2] similarly, for the training efficiency of multi device cooperative scheduling in integrated energy system, Xiang Zhou and others proposed a DRL framework based on asynchronous dominant actor critic (A3C), which can synchronously train the global/local strategy network and value network through CPU multithreading, so that the training convergence speed of ies system in Belgium grid is more than 40% higher than that of single thread dqn, and avoid the accumulation of error caused by discretization - the system operation cost after 30000 rounds of training stable convergence, only 0.33% higher than the traditional CPLEX

optimization, which verifies the effectiveness of multi-threaded parallel training in reducing the difficulty of DRL training [8].

- the differentiable holomorphic embedded power flow model ensures the operability of the solution based on the continuation power flow principle, and adaptively updates the Lagrange multiplier processing constraints combined with the constraint strategy optimization (CPO). The verification in IEEE 5/30/118/300 node system shows that the training convergence speed of the algorithm is much faster than that of sac, PPO and other mainstream DRLs, and the calculation speed is up to 113 times higher than that of the commercial nonlinear solver when running online (the calculation time of 30 node system is only 0.102s) [3].

Data enhancement solves the problem of insufficient training data, while physical informed fusion reduces ineffective exploration through prior knowledge. The fault scenario generated by wcgan-gp increases the amount of training data by one order of magnitude, and the physical constraint of pinn increases the convergence speed of the model by 18.7%~40%, significantly reducing the training threshold of traditional DRL [2,4].

2.4 Limited Generalization Ability

The performance of traditional DRL model drops sharply when the grid topology and res permeability change, and the generalization is poor. Existing studies have improved the model's Cross scene adaptability through spatio-temporal feature extraction and transfer learning technology:

- topological feature extraction enhances spatial generalization. Zhou et al. Added the graph neural network (GNN) layer to h-mappo, input the grid adjacency matrix and line admittance, and extract topological correlation features. The convergence rate in high load fluctuation scenarios (50%~150%) remains 100%, which is better than 68.8% of Matpower [2].
- time series feature capture enhances dynamic generalization. Gao and others use LSTM to extract the intra day time series dependence of photovoltaic output and input it into the DRL strategy network. When photovoltaic permeability changes, the light rejection rate is still controlled below 5%, and the generalization error is less than 3% [8].
- knowledge migration reduces the cost of scene adaptation. Through knowledge regularization, the scheduling cost error of ke-mappo model increases by only 8%~12% after grid topology adjustment, which is more than 15% lower than that of no knowledge enhancement model; The future domain adaptive technology can achieve 95% performance of 5% new data [6].

Spatio temporal feature extraction enables DRL model to capture the laws of power grid topology association and time series fluctuation, and knowledge transfer realizes cross scene knowledge reuse. Gnn+drl still maintains full convergence under severe load fluctuations, and the topology adaptation error of knowledge enhancement model is reduced by more than 15%, which effectively solves the generalization problem of traditional DRL that "scene change requires retraining" [2,6].

3 Existing Technical Bottlenecks

3.1 Multidimensional Generalization Ability is Still Insufficient

The existing DRL models are mostly trained for specific systems (such as IEEE 33 node, rts-gmlc system), so when the grid topology and res permeability change, the models basically need to be retrained. For example, the DRL model performs best when the PV penetration rate is 30%, but when the penetration rate increases to 50%, the constraint satisfaction rate decreases to 75% [9]; Shi Wenlong's research shows that without the knowledge enhanced mappo model, the dispatching cost error increases by 8%~12% after the actual power grid topology adjustment, which indicates that the generalization ability of DRL is still not well solved [6].

3.2 Multi Time Scale Coordination Difficulty

OPF needs to coordinate multi time scale tasks such as "second level res fluctuation regulation" and "hour level unit output scheduling". The existing DRL is mostly optimized for a single time scale, and cross scale coordination is easy to lead to decision lag. Wu et al. Found in IEEE 118 bus system that without considering the multi time scale DRL model, the line overload risk increased by 15% in the peak valley switching period (such as 8 a.m. and 18 p.m.) [5].

3.3 High Dependence on Data Quality and Knowledge Acquisition

DRL training requires a large number of high-quality historical data, but the actual power grid data has problems such as noise and missing. 10% noise will increase the OPF cost error of DRL by 5%~8% [7]; At the same time, it is difficult to obtain tacit knowledge (such as dispatcher experience and extreme scenario response rules). Although ke-mappo model can embed some tacit knowledge, it still cannot cover all non intuitive decision logic [6,10].

3.4 It is Difficult to Balance the Safety and Efficiency of Large-scale Systems

In large-scale systems such as activesg2000, although the security constraint DRL (such as pd-ppo) can ensure constraint satisfaction, the decision-making time increases linearly with the number of nodes, and overly conservative strategies will lead to economic losses. Wu et al's research shows that although the security layer method can make the constraint satisfaction rate reach 100%, the operating cost is 18.7% higher than pdo-amg [5].

4 Future Development Directions

4.1 Integration of Multi Time Scale Hierarchical DRL and Knowledge Transfer

The hierarchical architecture is designed. The upper DRL handles long-time scale unit scheduling (embedding long-term operation knowledge), while the lower DRL handles short-time scale res fluctuation regulation (embedding real-time safety rules), and

migrates the trained model knowledge to a new scene through migration learning, reducing the need for retraining data.

4.2 DRL Driven by Physics Data Knowledge Triple Fusion

Deeply integrate the physical model, data-driven model and domain knowledge (scheduling rules, weak branch information), integrate the graph neural network (GNN), embed the dynamic changes of power grid topology, and improve the adaptability to grid reconfiguration, and improve the interpretability and robustness of the model [3, 10].

4.3 Human Machine Hybrid Enhanced Intelligence

Combined with the dispatcher's experience and DRL decision-making, the human-computer interaction interface is designed - DRL provides the basic scheduling scheme. The dispatcher forms a closed loop of "machine decision-making human correction model feedback" by modifying the rule optimization scheme, so as to improve the reliability of decision-making in complex scenes. Shi Wenlong's practice in Tianjin power grid shows that the success rate of man-machine hybrid decision-making is 12% higher than that of pure DRL decision-making in extreme scenarios [6].

5 Conclusion

Deep reinforcement learning has been well applied in operation control (voltage regulation, frequency control, etc.), emergency control (fault induced delayed voltage recovery (fidvr), power system emergency frequency control (psefc)) and small signal stability control (suppression of low frequency oscillation (LFO), ultra low frequency oscillation (ulfo)), and also provides an efficient solution for the optimal power flow problem of existing renewable energy power systems.From the small-scale decision-making of single agent to the large-scale collaboration of multi-agent, from the feasibility of physical constraint fusion to GNN to improve the space-time perception ability, and then to the secure DRL and data knowledge double enhancement technology, this paper summarizes the solutions of DRL in the OPF problem of high proportion of renewable energy connected to the power system. If the four defects mentioned in this paper can be solved, it can promote its application in the actual power system.

Therefore, future research should combine these human-computer cooperation technologies such as transfer learning, physical data knowledge integration, and lightweight design to further enhance the practicability of DRL in complex power systems and provide corresponding support for the safe and economic operation of power systems with a high proportion of renewable energy access.

References

1. Aien, M., Hajebrahimi, A., Fotuhi-Firuzabad, M.: A comprehensive review on uncertainty modeling techniques in power system studies. Renewable and Sustain. Energy Rev. 57, 1077–1089 (2016)

2. Zhou, L., Huo, L., Liu, L., et al.: Optimal Power Flow for High Spatial and Temporal Resolution Power Systems with High Renewable Energy Penetration Using Multi-Agent Deep Reinforcement Learning. Energies 18(7), 1809 (2025)
3. Ahmed, S, Khaled, A. J., Xian, Z., Hatem, Z., Ahmed, A., G, Wang., E, E.: Efficient optimal power flow learning: A deep reinforcement learning with physics-driven critic model (2025)
4. Tudoras-Miravet, À., González-Iakl, E., Gomis-Bellmunt, O., et al.: Physics-Informed Neural Networks for Power Systems Warm-Start Optimization. IEEE Access 12, 135913–135928 (2024)
5. Wu, P., Chen, C., Lai, D., et al.: Real-Time Optimal Power Flow Method via Safe Deep Reinforcement Learning Based on Primal-Dual and Prior Knowledge Guidance. IEEE Transactions on Power Systems 40(1), 597–609 (2025)
6. Shi, W. L.: Data and Knowledge Enhanced Multi-Agent Deep Reinforcement Learning for Real-Time Power Grid Scheduling. (2024)
7. Dai, Y., Xu, W., Yan, M., et al.: Deep Reinforcement Learning Explanation-Assisted Integer Variable Reduction Method for Security-Constrained Unit Commitment. Engineering Applications of Artificial Intelligence 144, 110139 (2025)
8. Zhou, X., Wang, J., Wang, X., et al.: Optimal Dispatch of Integrated Energy System Based on Deep Reinforcement Learning. Energy Reports 9, 373–378 (2023)
9. Wang, R., Gao, H., Luo, L., et al.: Review of Research on New Distribution System Optimization Operation Based on Deep Reinforcement Learning. Electric Power Automation Equipment 45(9), 1–15 (2025)
10. Li, Q., Lin, T., Yu, Q., et al.: Review of Deep Reinforcement Learning and Its Application in Modern Renewable Power System Control. Energies 16(10), 4143 (2023)