A Model-Free Approach for Dynamic Operating Envelopes Calculation Based on Multi-agent Reinforcement Learning

1st Yuchen Qi
Shenzhen International Graduate School
Tsinghua University
Shenzhen, China
qiyc24@mails.tsinghua.edu.cn

2nd Ye Guo
Department of Building Environment and Energy Engineering
Hong Kong Polytechnic University
Kowloon, Hong Kong
ye.guo@polyu.edu.hk

3rd Yinliang Xu
Shenzhen International Graduate School
Tsinghua University
Shenzhen, China
xu.yinliang@sz.tsinghua.edu.cn

4th Zhenjia Lin
Department of Building Environment and Energy Engineering
Hong Kong Polytechnic University
Kowloon, Hong Kong
epjack.lin@polyu.edu.hk

***Abstract*—The widespread adoption of Distributed Energy Resources (DERs) has introduced new challenges in power system optimization. Traditional Dynamic Operating Envelopes (DOEs) rely on electrical models, but these face issues with data scarcity and computational complexity when models are unavailable. This paper proposes a model-free DOE allocation method based on deep learning and reinforcement learning. A deep learning model is trained to map active and reactive power to voltage magnitudes, replacing traditional power flow calculations. A multi-agent reinforcement learning algorithm is then used to optimize DOE allocation by training DER and DSO agents. Simulation results on 33-bus and 141-bus system show that the model-free method allocates DOEs efficiently while maintaining system safety, with errors kept within 15% compared to model-based methods in 33-bus system. This approach provides a novel, data-driven solution for intelligent power system optimization.**

***Index Terms*—Convolutional neural networks, Dynamic operating envelopes, Multi-agent reinforcement learning**

I. INTRODUCTION

The rapid adoption of distributed energy resources (DERs) and the decreasing costs of solar photovoltaics and battery storage have significantly changed power system operations. Traditionally, power systems assumed unidirectional electricity flow from large power plants to consumers. However, with DERs, electricity now flows bidirectionally, as consumers can also generate and inject excess energy back into the grid. This shift introduces challenges in grid operation, particularly in low-voltage distribution networks, where DER integration can cause voltage violations [1].

To prevent voltage violations, Distribution Network Operators (DSOs) often impose fixed power output limits on DER buses. However, this approach does not account for dynamic market demands and generation patterns, limiting aggregator participation and causing energy waste. Dynamic Operating Envelopes (DOEs) have been proposed as an alternative, defining time- and location-varying power limits based on system capacity to ensure efficient DER integration while maintaining voltage safety [2].

Conventionally, the calculation of Dynamic Operating Envelopes (DOE) relies heavily on Optimal Power Flow (OPF) frameworks. These approaches necessitate detailed grid topology models, precise user demand data, and accurate generation forecasts [3]–[6]. However, the practical applicability of these methods is often hindered by the inherent difficulty in obtaining and maintaining accurate grid parameters [7].

To circumvent reliance on explicit grid models, recent research has pivoted toward data-driven or model-free paradigms. For instance, Bassi et al. [8] trained a deep learning model to serve as a surrogate for power flow calculations, employing heuristic algorithms for DOE allocation. Similarly, Cao et al. [9] proposed a model-free voltage regulation scheme for three-phase unbalanced networks, integrating surrogate modeling with Deep Reinforcement Learning to mitigate rapid voltage fluctuations caused by photovoltaic variability. Furthermore, Kumarawadu et al. [10] utilized smart meter data

The work was supported in part by the National Natural Science Foundation of China under Grant 52377105. (Corresponding author: Ye Guo, e-mail: ye.guo@polyu.edu.hk)

979-8-3315-8138-1

to estimate voltage sensitivity coefficients, effectively approximating the electrical model to facilitate OPF-based calculations.

Despite these advancements, significant gaps remain. Most existing data-driven approaches neglect the heterogeneity of DERs, failing to account for their distinct generation capacities and willingness to participate. Crucially, these methods typically overlook the game-theoretic interactions among DERs and the imperative of fairness in allocation. Moreover, the reliance on centralized data aggregation in prior studies raises privacy concerns, as it necessitates the disclosure of detailed DER operational information.

To address the aforementioned challenges, a novel model-free framework for DOE calculation is proposed. The key contributions of this paper are summarized as follows:

- A hybrid data-driven framework integrating deep learning with Multi-Agent Reinforcement Learning (MARL) is developed, which eliminates the dependency on explicit grid topology and parameters.
- A decentralized DOE allocation strategy featuring specialized reward functions and action spaces for DSO and DER agents is designed. This mechanism explicitly accounts for the generation capacities and strategic bidding behaviors of DERs. By modeling the interactions as a game-theoretic process, the proposed strategy enhances fairness and resource utilization while preserving the information privacy of individual participants.

II. Framework

The model-free approach to DOE calculation addresses two key challenges: first, determining system safety by calculating bus voltages under known conditions, and second, allocating DOEs effectively. The framework proposed in this paper is shown in Fig. 1.

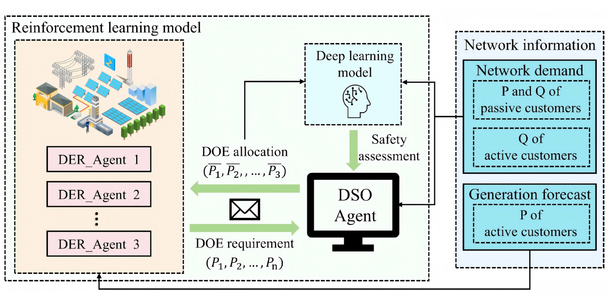

Fig. 1. Framework for model-free DOE calculation.

First, a deep learning model is pre-trained to map active and reactive power inputs to bus voltage magnitudes, serving as a high-fidelity surrogate for traditional power flow analysis. By decoupling the voltage prediction from the policy optimization, the deep learning model provides near-instantaneous safety feedback, which significantly reduces the computational overhead and improves the convergence stability of the reinforcement learning training process. Consequently, this hybrid approach leverages the predictive strength of deep learning to establish a reliable safety criterion for assessing voltage limit violations during the subsequent DOE allocation phase.

Next, we train a reinforcement learning model to allocate DOEs. This process employs a multi-agent reinforcement learning strategy. Each DER bus is treated as an agent that learns how to maximize its own benefit by requesting more DOE allocation. In each learning round, the agents only have access to the maximum power they can inject, and they report their DOE requirements based on their current state.

To capture the interaction between DER buses and the DSO, the training process allows DER buses to request more DOE than their maximum available power in order to secure additional allocation. However, this comes with the risk of being penalized if the reported DOE exceeds their capacity. Meanwhile, the DSO is also an agent, aiming to learn a reasonable DOE allocation strategy. The DSO's goal is to maximize the total DOE distribution while ensuring system safety and fairness. During each learning round, the DSO is aware of the active and reactive power of all non-DER buses, the reactive power of DER buses, and the DOE requests made by DER buses. Using this information, the DSO allocates DOEs and uses the pre-trained deep learning model to verify if the voltage at each bus remains within safe limits. Any violation results in a penalty.

Once the reinforcement learning model is trained, it can take inputs of active and reactive power from non-DER buses, as well as the reactive power and maximum injection power of DER buses. The model then outputs the optimal DOE allocation for each DER bus, providing a model-free approach to DOE calculation.

III. Methodology

A. Deep Learning Model Generation

In this study, we use a deep learning approach to replace power flow calculations without relying on an electrical model. The goal is to train a model that takes the active and reactive power at each bus as input and outputs the corresponding bus voltage magnitude, which will be used for subsequent safety assessments. Specifically, we employ a convolutional neural network (CNN) [11] augmented with an attention mechanism [12] to construct the deep learning model. The CNN module effectively extracts spatial features from the grid topology, while the attention mechanism captures complex, non-linear dependencies between specific buses.

979-8-3315-8138-1

The CNN operation is defined as:

$$x_{l+1} = \text{ReLU}(W_l * x_l + b_l), \tag{1}$$

where x_l represents the input feature map to the l-th layer, W_l is the convolution kernel, b_l is the bias term, and $*$ denotes the convolution operation. The ReLU function introduces non-linearity to the model.To prioritize critical features, the attention mechanism computes weights α_{ij}:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_k \exp(e_{ik})}, \tag{2}$$

where $\alpha_{ij} \in [0, 1]$ represents the normalized attention weight from the i-th input to the j-th feature, and e_{ij} is the similarity score between the query and the key vectors. This mechanism allows the model to allocate different levels of attention to various inputs, ensuring that important features are given priority in the prediction of bus voltage magnitudes.

After processing through the CNN and attention modules, the model's final output is passed through a fully connected layer for regression, yielding the predicted voltage magnitude for each bus.

B. Reinforcement Learning Model Generation

This section outlines the development of a multi-agent reinforcement learning model for DOE allocation. We formulate DOE allocation as a multi-agent problem where each DER bus acts as an agent. Agents report DOE requests based on capacity ($p_i^{\max}$) but may exaggerate to secure higher allocations. The i-th DER agent's reward balances allocation utility, over-reporting penalties, and risk:

$$r_i = \alpha \cdot a_i - \beta \cdot \max(0, a_i - p_i^{\max}) - \gamma \cdot \rho_i, \tag{3}$$

where a_i is the allocated DOE. Coefficients $\alpha, \beta, \gamma > 0$ weigh the terms. The risk factor ρ_i tracks exaggeration behavior:

$$\rho_i^{(t)} = \lambda \cdot \rho_i^{(t-1)} + \mathbb{I}(a_i > p_i^{\max}) \cdot \frac{a_i - p_i^{\max}}{p_i^{\max}}, \tag{4}$$

where $\lambda \in [0, 1]$ is the decay factor, representing the retention rate of previous risks on the current state and $\mathbb{I}(\cdot)$ is the indicator function. The risk factor increases when the DER agent exaggerates its DOE request and decays otherwise. This mechanism discourages excessive distortion.

The DSO, as another agent, optimizes a global objective comprising total allocation, fairness, and safety:

$$r^{DSO} = \eta \left(\sum_i a_i \right) + \varphi \cdot \frac{J(a) + J(\bar{a})}{2} + \psi \cdot \tilde{v}, \tag{5}$$

where $\eta, \varphi > 0$ are weighting coefficients for total allocation and fairness, while $\psi < 0$ is the penalty coefficient for voltage violations. a_i represents the DOE allocated to the i-th DER bus. Fairness is quantified by the Jain index $J(a) \in [1/n, 1]$ [13]:

$$J(a) = \frac{\left(\sum_i a_i\right)^2}{n \sum_i a_i^2 + 10^{-6}}, \tag{6}$$

where n is the number of DERs. $\bar{a}$ denotes the moving average over window $W \in \mathbb{Z}^+$:

$$\bar{a} = \frac{1}{W} \sum_{k=t-W+1}^{t} a^{(k)}. \tag{7}$$

The third term addresses voltage limit violations. $\tilde{v}$ represents the cumulative penalty for all voltage violations across the network. When any bus's voltage exceeds 1.05 p.u., a penalty is applied:

$$\tilde{v} = \sum_i \left[\exp\left(10^3 \times (v_i - 1.05)\right) - 1 \right] \cdot \mathbf{1}_{\{v_i > 1.05\}}, \tag{8}$$

where v_i is the voltage magnitude at the i-th bus predicted by the deep learning model, and $\mathbf{1}_{\{v_i > 1.05\}}$ is an indicator function. The constant 10^3 serves as a scaling factor to sharply penalize violations.

To train the DER and DSO agents, we employ the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm [14]. MADDPG adapts the DDPG algorithm to multi-agent settings. The action is determined by the following equation:

$$\mu_{\theta_i}(s_i) = \text{clip}\left(\pi_{\theta_i}(s_i) + \mathcal{N}_i, -1, 1\right), \tag{9}$$

where $\mu_{\theta_i}(s_i)$ is the action chosen by the i-th agent at state s_i, and $\mathcal{N}_i$ is the exploration noise. The critic network for each agent is updated by minimizing the Bellman error:

$$\begin{aligned} L_i = \mathbb{E}_{(s_i, a_i, r_i, s_i') \sim D} \Bigg[\Big(& Q_{\theta_i}(s_i, a_i) - \\ & \left(r_i + \gamma \cdot Q_{\theta_i'}(s_i', a_i') \right) \Big) \Bigg]^2, \end{aligned} \tag{10}$$

where $Q_{\theta_i}(s_i, a_i)$ is the critic's Q-value, and $\gamma \in [0, 1)$ is the discount factor. The actor network is updated by maximizing the critic's Q-value:

$$\nabla_{\theta_i} J_{\text{actor}} = \mathbb{E}_{s_i \sim D} \left[\nabla_{a_i} Q_{\theta_i}(s_i, a_i) \cdot \nabla_{\theta_i} \mu_{\theta_i}(s_i) \right]. \tag{11}$$

Finally, target networks are updated using a soft update method:

$$\theta_i' \leftarrow \tau \theta_i + (1 - \tau) \theta_i', \tag{12}$$

where $\tau \in (0, 1)$ is the soft update parameter.

MADDPG enables efficient joint training of DER and DSO agents, balancing exploration and exploitation while ensuring system stability and fairness in DOE allocation.

IV. Case Study

A. Dataset Generation

Simulations were performed on the 33-bus and 141-bus systems. The datasets were generated using MATPOWER's `case33bw.m` [15] and `case141.m` [16], with load data scaled based on smart meter data from [17]. 10% of the buses were designated as DER buses (4 in the 33-bus system and 14 in the 141-bus system), with their generation scaled according to [18]. The dataset includes active/reactive power and voltage magnitudes. A total of 5,000 data points were used for the deep learning model, and 10,000 for reinforcement learning.

B. Results

Figure 2 shows the deep learning model's performance, using the output voltage at bus 5 for both systems. The deep learning model achieved R^2 values of 0.9997 and 0.9999, and RMSE values of 0.0004 and 0.0001 for the 33-bus and 141-bus systems, respectively. The regression lines closely align with $y = x$, demonstrating the model's accuracy and adaptability across systems.

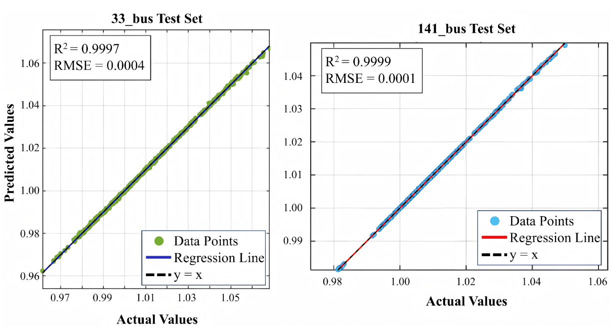

Fig. 2. Voltage magnitude calculation performance at bus 5 in 33-bus and 141-bus systems.

The trained deep learning model was used for safety assessment, guiding the reinforcement learning model's training. The reward curve from reinforcement learning is shown in Fig. 3. In the 33-bus system, both the DER and DSO agent reward curves showed significant improvement after 220 episodes. The DER agent's reward stabilized, while the DSO agent's reward decreased slightly before stabilizing. In the 141-bus system, the DER and DSO agent rewards started improving after around 100 episodes, with the DER agent stabilizing and the DSO agent exhibiting minor fluctuations before stabilizing. This indicates that the reinforcement learning model successfully identified the optimal strategy and achieved stable convergence in both the 33-bus and 141-bus systems.

Table I compares the proposed method with MAPPO and MASAC models. Although MAPPO and MASAC show higher DOE allocations, they also exhibit higher violation magnitudes, voltage violation rates, and lower fairness values. This is due to their variance in sampling near safety boundaries, sacrificing safety for extreme allocations. In contrast, the proposed method ensures better safety and fairness by using a deterministic policy, providing consistent safe actions once training converges.

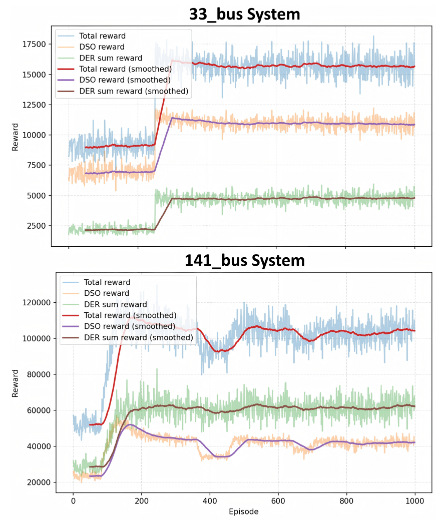

Fig. 3. The reward curve from the multi-agent reinforcement learning model training.

To validate the method's effectiveness, taking the 33-bus system as an example, the reinforcement learning model was used for DOE allocation, shown in Fig. 4. Panel (a) compares the system's maximum voltage before and after model-free DOE allocation. Before allocation, the voltage was calculated assuming all DER buses operated at maximum net injection power. After allocation, the voltage decreased at most limit-exceeding points, with only a few time points exceeding 1.05, indicating the safety of the method. At time points where the voltage was below the threshold before allocation, the system's voltage remained unchanged, demonstrating effective use of resources.

Panels (b) and (c) of Fig. 4 compare model-free and model-based DOE allocations. The model-free method kept the error in DOE allocation for most DER buses within 15% of the model-based approach, confirming its effectiveness.

V. Conclusion

This paper develops a model-free DOE allocation framework that effectively circumvents the need for explicit grid models by integrating a deep learning surrogate with a multi-agent reinforcement learning strategy. Validated on both 33-bus and 141-bus systems, the proposed CNN-based model achieves near-

979-8-3315-8138-1

TABLE I
COMPARISON BETWEEN DIFFERENT MODELS FOR 33-BUS AND 141-BUS SYSTEMS.

	33-bus system			141-bus system		
	Proposed	MAPPO	MASAC	Proposed	MAPPO	MASAC
Avg. Violation (10^{-3} p.u.)	2.568	6.883	6.500	9.389	644.4	297.3
Violation Rate (%)	0.585	3.081	2.750	1.132	14.25	10.08
Avg. DOE Allocation (MW)	6.123	9.767	9.541	138.7	268.2	207.56
Fairness (Jain Index)	0.8426	0.6938	0.6954	0.9893	0.9573	0.9642

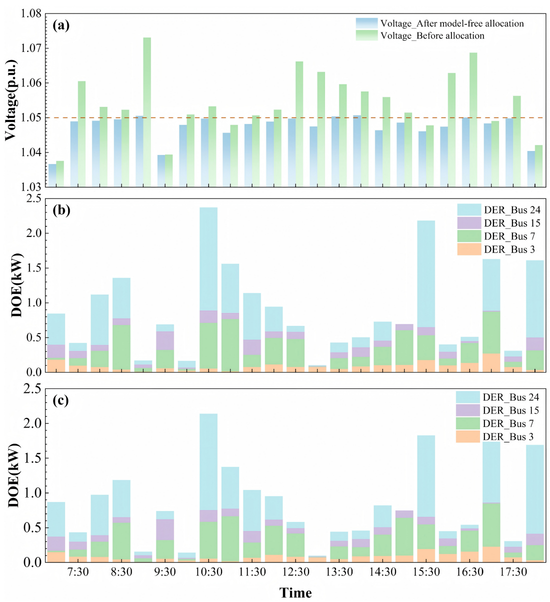

Fig. 4. DOE allocation with different methods in 33-bus system. (a) Maximum voltage magnitudes before and after DOE allocation. (b) DOE allocation using the model-free method. (c) DOE allocation using the model-based method.

instantaneous voltage estimation with high fidelity. Crucially, the MADDPG-based allocation strategy ensures superior system safety and fairness compared to MAPPO and MASAC. During model validation, the model-free approach effectively reduced the system's maximum voltage to within the safety threshold. Compared to model-based methods, the error in DOE allocation remained within 15%, ensuring both voltage safety and fair distribution.

REFERENCES

[1] K. Petrou, A. T. Procopiou, L. F. Ochoa, T. Langstaff, and J. Theunissen, "Impacts of price-led operation of residential storage on distribution networks: An australian case study," in *2019 IEEE Milan PowerTech*. IEEE, 2019, pp. 1–6.

[2] L. Blackhall, "On the calculation and use of dynamic operating envelopes," *Australian National Univerity, Canberra, ACT*, 2020.

[3] Z. Jiang, Y. Guo, and J. Wang, "Dynamic operating envelopes embedded peer-to-peer-to-grid energy trading," *Applied Energy*, vol. 377, p. 124554, 2025.

[4] Z. Jiang and Y. Guo, "Bargaining-based approach for dynamic operating envelope allocation in distribution networks," *IEEE Transactions on Smart Grid*, 2025.

[5] B. Liu and J. H. Braslavsky, "Robust dynamic operating envelopes via superellipsoid-based convex optimisation in unbalanced distribution networks," *IEEE Transactions on Power Systems*, vol. 39, no. 2, pp. 4775–4778, 2024.

[6] B. Liu, J. H. Braslavsky, and N. Mahdavi, "Linear opf-based robust dynamic operating envelopes with uncertainties in unbalanced distribution networks," *Journal of Modern Power Systems and Clean Energy*, vol. 12, no. 4, pp. 1320–1326, 2023.

[7] F. Geth, M. Vanin, and D. Hertem, "Data quality challenges in existing distribution network datasets," 2023.

[8] V. Bassi, L. F. Ochoa, T. Alpcan, C. Leckie, and M. Z. Liu, "Smart meter data and operating envelopes in lv networks: A model-free approach," in *2023 IEEE PES Innovative Smart Grid Technologies Latin America (ISGT-LA)*. IEEE, 2023, pp. 1–5.

[9] D. Cao, J. Zhao, W. Hu, F. Ding, N. Yu, Q. Huang, and Z. Chen, "Model-free voltage control of active distribution system with pvs using surrogate model-based deep reinforcement learning," *Applied Energy*, vol. 306, p. 117982, 2022.

[10] A. Kumarawadu, M. I. Azim, M. Khorasany, R. Razzaghi, and R. Heidari, "Smart meter data-driven dynamic operating envelopes for ders," *Applied Energy*, vol. 384, p. 125469, 2025.

[11] K. O'shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] A. B. Sediq, R. H. Gohary, R. Schoenen, and H. Yanikomeroglu, "Optimal tradeoff between sum-rate efficiency and jain's fairness index in resource allocation," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3496–3509, 2013.

[14] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in neural information processing systems*, vol. 30, 2017.

[15] H.-C. Chang and C.-C. Kuo, "Network reconfiguration in distribution systems using simulated annealing," *Electric Power Systems Research*, vol. 29, no. 3, pp. 227–238, 1994.

[16] H. Khodr, F. G. Olsina, P. M. De Oliveira-De Jesus, and J. Yusta, "Maximum savings approach for location and sizing of capacitors in distribution systems," *Electric power systems research*, vol. 78, no. 7, pp. 1192–1203, 2008.

[17] F. Bu, Y. Yuan, Z. Wang, K. Dehghanpour, and A. Kimber, "A time-series distribution test system based on real utility data," in *2019 North American Power Symposium (NAPS)*. IEEE, 2019, pp. 1–6.

[18] Y. Chen and J. Xu, "Solar and wind power data from the chinese state grid renewable energy generation forecasting competition," *Scientific data*, vol. 9, no. 1, p. 577, 2022.

979-8-3315-8138-1