Safe Reinforcement Learning for 110kV Distribution Grid Restoration (Draft)

Prospective Master Project-Thesis Draft
Changhun Kim[1*] Simon Linnert [2]
[1]*Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nuremberg*
[2]*Institute of Electrical Energy Systems, Friedrich-Alexander-University Erlangen-Nuremberg*
Erlangen, Germany

***Abstract*—Restoring power after a major outage (Netzwiederaufbau) in a 110kV distribution network is a complex sequential decision-making task. Operators must coordinate switching operations, generator starts, and load reconnections under tight safety constraints (voltage, frequency, equipment limits). Traditional rule-based restoration plans struggle to adapt to unforeseen scenarios, motivating the use of reinforcement learning (RL) to learn adaptive strategies. Recent studies have shown that deep RL can successfully orchestrate distribution system restoration using distributed energy resources and network switching. However, applying RL to power system restoration requires addressing safety and trust. An inappropriate action (e.g. closing a breaker out of sequence) can jeopardize system stability, so the learning process and the learned policy must strictly respect operational constraints. Moreover, a black-box RL agent's suggestions may not be trusted by human operators due to lack of transparency.**

***Index Terms*—We aim to develop a safe, real-time-capable RL methodology that assists distribution-network operators during restoration at $110\,\mathrm{kV}$. The policy must (i) interface with DIgSILENT PowerFactory's real-time RMS simulator, (ii) yield millisecond inference for operator decision support, (iii) guarantee operational safety via constrained or shielded learning, and (iv) provide interpretable rationales. confidence.**

I. Methodology Overview

We propose a two-stage deep-RL framework: (1) supervised imitation of *Restoration Typicals*, followed by (2) safe RL optimisation.

A. Environment Setup

The $110\,\mathrm{kV}$ network dynamics are modelled in DIgSILENT PowerFactory (RMS). At each RL step the agent selects a *Restoration action* (e.g. close tie switch, energise line, start generator); the simulator then resolves transients and returns the next state and reward. The simulator can be accelerated off-line for training, whereas the trained policy runs in real time for operator assistance.

B. State Representation

Observations include bus voltages, frequencies, breaker states, available generation and load levels. To capture topology, we embed the grid as a graph and apply a graph neural network (GNN) encoder [4].

C. Action Space: Restoration Typicals

Rather than arbitrary controls, we expose a discrete set of expert-derived *Restoration Typicals*. Constraining the agent to these feasible high-level actions reduces search space and prefilters unsafe moves, supplying a safety prior from the outset.

D. Reward Design

We reward load restored (MW or priority loads served) and penalise elapsed time and every constraint violation (out-of-range voltage, line overload, frequency excursion). Episodes terminate with a large negative reward if protective relays would trip [1]. A small per-switch penalty discourages excessive operations [5].

II. Stage 1: Supervised Imitation Learning

Historical expert trajectories and offline-optimised Typicals furnish state–action pairs. We train a policy network via behaviour cloning:

$$\mathcal{L}_{\mathrm{BC}} = \mathbb{E}_{(s,a)\sim\mathcal{D}}[-\log \pi_\theta(a \mid s)]\,. \tag{1}$$

The resulting policy π_0 replicates established restoration sequences with high fidelity, providing a safe initialisation and an *action filter* for Stage 2 [6].

III. Stage 2: Safe RL Optimisation

A. Constrained PPO

We refine π_0 using *Lagrangian Proximal Policy Optimisation* (PPO-L) [7]. Grid restoration is formulated as a constrained Markov decision process (CMDP):

$$\begin{aligned} \max_\pi J_r(\pi) &= \mathbb{E}\Big[\textstyle\sum_t \gamma^t r_t\Big] \\ \text{s.t. } J_c(\pi) &= \mathbb{E}\Big[\textstyle\sum_t \gamma^t c_t\Big] \le \epsilon, \end{aligned} \tag{2}$$

where c_t counts safety violations. PPO-L updates minimise the clipped surrogate while adjusting a Lagrange multiplier λ to enforce $J_c \le \epsilon$.

B. Shielded Action Selection

A runtime *shield* intercepts unsafe actions before execution, replacing them with the nearest safe alternative [8]. Shield interventions are logged and penalised, steering the policy to self-compliance.

IV. Integration and Deployment

The trained policy is exported (ONNX) and embedded in a Python-based decision-support tool interfaced with PowerFactory. Inference latency is $\approx 3\,\mathrm{ms}$, well below SCADA refresh cycles. Operators receive top-N ranked suggestions, each accompanied by rule-based and XAI explanations (decision-tree surrogate, SHAP) [9].

V. Conclusion

Our two-stage framework combines expert imitation, constrained PPO and shielding to deliver **safe, interpretable, real-time** restoration guidance for $110\,\mathrm{kV}$ grids. By grounding decisions in Restoration Typicals and guaranteeing constraint adherence, the approach accelerates recovery while maintaining operator trust.

References

[1] X. Chen, Y. Xu, and P. Zhang, "Deep reinforcement learning for distribution system restoration with DER coordination," *IEEE T. Smart Grid*, vol. 13, no. 2, pp. 987–999, 2022.

[2] J. Ding, H. Wang, and S. Low, "Safe policy gradient for microgrid black-start restoration," in *Proc. IEEE PES GM*, 2024.

[3] H. Liu *et al.*, "Explainable reinforcement learning: A survey," *ACM Comput. Surveys*, vol. 55, no. 7, pp. 1–38, 2023.

[4] R. Li, T. Liu, J. Yu *et al.*, "Graph neural network based voltage-control reinforcement learning for distribution systems," *IEEE T. Smart Grid*, vol. 12, no. 6, pp. 5269–5280, 2021.

[5] A. Molina García *et al.*, "Switching impact on MV equipment—a six-year field study," *CIRED Workshop*, 2020.

[6] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *AISTATS*, 2011.

[7] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proc. ICML*, 2017.

[8] M. Alshiekh, R. Bloem, R. Ehlers *et al.*, "Safe reinforcement learning via shielding," in *AAAI*, 2018.

[9] M. Du, N. Liu, Q. Hu *et al.*, "Techniques for interpretable deep learning," *Commun. ACM*, vol. 63, no. 1, pp. 68–77, 2020.

Acknowledgment

References