Multi-Agent Reinforcement Learning for Real-Time Frequency Regulation in Power Grids: Final Report

Derek Smith, Matthew Vu
ES 158: Sequential Decision Making in Dynamic Environments
Harvard University

December 14, 2025

Abstract

We implement Multi-Agent Proximal Policy Optimization (MAPPO) for real-time frequency regulation in a simulated 20-bus power grid with 10 heterogeneous agents (2 batteries, 5 gas generators, 3 demand response units). Our centralized training, decentralized execution (CTDE) approach addresses a cooperative Multi-Agent MDP with continuous state/action spaces, partial observability, and safety constraints. Through systematic debugging—addressing reward scaling, capacity mismatch, observation normalization, and curriculum learning—we achieved stable value function learning with critic loss converging from ∼29 to ∼22. However, policy performance exhibited the characteristic "forgetting" phenomenon of multi-agent systems: reward peaked at ∼165 around episode 1000 before degrading to ∼120 by episode 2000. We analyze this coordination collapse, attributing it to non-stationarity, credit assignment difficulties, and the exponential complexity of coordinating 10 independent agents under partial observability. Our results highlight both the promise and fundamental challenges of applying deep MARL to safety-critical infrastructure control.

1 Introduction

Modern power grids with $> 30\%$ renewable penetration face frequency stability challenges due to reduced inertia and stochastic generation, causing \$10B+ annual regulation costs [4]. Traditional PI controllers and MPC struggle to scale with grid complexity [3, 7]. Multi-agent reinforcement learning offers coordinated, adaptive control with potential 20–40% cost reduction [8].

We explore **Centralized Training, Decentralized Execution (CTDE)** using MAPPO, where a centralized critic observes global state during training, but agents execute using only local observations. We formulate frequency regulation as a cooperative MA-MDP with $N = 10$ heterogeneous agents (2 batteries, 5 gas plants, 3 demand response units) across a 20-bus network governed by swing equation dynamics.

Challenges: Continuous spaces ($\mathcal{S} \subseteq \mathbb{R}^{55}$, $\mathcal{O}^i \in \mathbb{R}^{15}$), partial observability, stochastic disturbances, hard safety constraints (± 1.5 Hz), and multi-agent coordination (non-stationarity, credit assignment).

Contributions: (1) 20-bus power grid simulator with realistic dynamics; (2) MAPPO implementation with CTDE; (3) systematic debugging methodology for reward scaling, capacity matching, and curriculum learning; (4) analysis of multi-agent coordination challenges with experimental results.

2 Environment Design

2.1 Grid Topology and Agents

We implement a 20-bus power network with 10 heterogeneous agents: 2 batteries (50 MW/min, [0,100] MW), 5 gas plants (10 MW/min, [50,500] MW), and 3 demand response units (5 MW/min, [-200,0] MW). Seven renewable sources provide 50–300 MW each with stochastic variation.

Capacity matching is critical: total controllable generation (∼3300 MW) must exceed load range. We set loads to [1500, 3000] MW to ensure agents can always balance the grid—without this, learning is impossible.

2.2 Physics and Dynamics

Frequency dynamics follow the **swing equation**:

$$\frac{df_k}{dt} = \frac{P_{\mathrm{mech},k} - P_{\mathrm{elec},k}}{2H_k \cdot S_{\mathrm{base}}} \tag{1}$$

with inertia $H_k \in [2, 7]$s, base power $S_{\mathrm{base}} = 10,000$ MVA, time step $\Delta t = 2$s (SCADA delay), and N-1 contingencies ($p = 0.001$/step).

2.3 State and Action Spaces

Local observation (15-dim): frequency deviation, load, own output, system frequency deviation, 5 neighbor frequencies, 3-step renewable forecast, time features, capacity utilization.

Global state (55-dim): All bus frequencies, generator outputs, renewables, loads, time features.

Actions: Continuous $[-1, 1]$ scaled by agent-specific ramp rates.

2.4 Curriculum Learning

We progressively tighten frequency bounds: Stage 1 (ep 0–1500): ±2.5 Hz; Stage 2: ±2.2 Hz; Stage 3: ±2.0 Hz; Stage 4 (ep 3500+): ±1.8 Hz. This allows agents to learn basic control before facing strict constraints.

3 Method: MAPPO

We selected MAPPO over MADDPG for stability (clipped objective), empirical performance on cooperative benchmarks [9], and hyperparameter robustness.

3.1 Architecture

Actor $\pi(a|o;\theta)$: $[15 \to 128 \to 128 \to 1]$ with LayerNorm/ReLU, outputs Gaussian $\mathcal{N}(\mu, \sigma^2)$, ∼19K parameters shared across agents.

Critic $V(s;\phi)$: $[55 \to 256 \to 256 \to 1]$ with LayerNorm/ReLU, takes global state, ∼81K parameters.

During training, critic accesses full state; during execution, actors use only local observations (CTDE).

3.2 Training Algorithm

Collect rollouts until buffer size $B = 2048$. Compute advantages via GAE-λ:

$$A_t = \sum_{l=0}^{T-t} (\gamma\lambda)^l \delta_{t+l}, \quad \delta_t = R_t + \gamma V(s_{t+1}) - V(s_t) \tag{2}$$

Update policy via PPO clipped objective over 10 epochs:

$$L^{\text{CLIP}} = \mathbb{E}\left[\min\left(r_t A_t, \text{clip}(r_t, 1-\epsilon, 1+\epsilon) A_t\right)\right] + \beta_{\text{ent}} H(\pi) \tag{3}$$

where $r_t = \pi_\theta(a_t|o_t)/\pi_{\theta_{\text{old}}}(a_t|o_t)$. Update critic via MSE: $L^{\text{CRITIC}} = \mathbb{E}[(V(s_t) - G_t)^2]$.

3.3 Hyperparameters

Actor LR	3×10^{-4}	Critic LR	1×10^{-3}
GAE λ	0.95	Discount γ	0.99
Entropy coef	0.02	Value coef	0.5
Clip ϵ	0.2	Grad norm	0.5
Buffer	2048	Batch	256

Implementation: ~1500 lines across `power_grid_env.py` (775 lines), `networks.py`, `mappo.py`, `buffer.py`, `train.py` with TensorBoard logging.

4 Training Challenges & Solutions

4.1 Initial Problems and Fixes

Early training exhibited severe instability: critic loss exploding to 10^{13}, episodes terminating at ~50 steps, and rewards in millions (negative). We systematically debugged these issues:

Problem	**Solution**
Exploding critic loss	Scaled rewards by $\div 100,000$
Immediate termination	Curriculum learning with relaxed bounds
Agents doing nothing	Added survival + stability bonuses
Capacity mismatch	Reduced load range to [1500, 3000] MW
Poor frequency signal	Normalized to frequency *deviations*

4.2 Key Fix: Capacity Mismatch

The most critical issue: with load range [2000, 5000] MW but only ~3300 MW controllable generation, agents *could not* balance the grid at high loads. Reducing to [1500, 3000] MW made learning feasible.

4.3 Reward Function

We redesigned rewards to align with curriculum bounds:

$$\begin{aligned} R_t = &-2000 \textstyle\sum_k (f_k - 60)^2 - 1000 \sum_k (\exp(2 \cdot \text{approach}_k) - 1) \\ &- \textstyle\sum_i C_i |a_t^i| - 0.05 \sum_i W_i (a_t^i)^2 + 5000 \cdot n_{\text{stable}} + 5000 \end{aligned} \tag{4}$$

where approach_k measures proximity to the *current curriculum* termination bound, not fixed ± 0.5 Hz.

4.4 Lessons Learned

1. **Ensure physical feasibility** before training—verify optimal policy can achieve the goal
2. **Reward scaling is critical**—normalize to $[-10, +10]$ range
3. **Align reward with termination**—if curriculum changes bounds, rewards must track
4. **Normalize observations**—use deviations, not raw values

These fixes transformed training from divergent to convergent. However, as Section 5 shows, deeper multi-agent coordination challenges remain.

5 Experiments & Results

5.1 Setup

Training: 2000 episodes, max 500 steps, buffer 2048, batch 256, 10 PPO epochs. Environment: load [1500, 3000] MW, 7 renewable sources, N-1 contingencies ($p = 0.001$), 2s SCADA delay, 4-stage curriculum.

5.2 Training Results

Figure 1 shows three phases: (1) rapid learning (ep 0–500, reward ∼130→150); (2) peak performance (ep 500–1200, reward ∼165); (3) degradation (ep 1200–2000, reward →120). Critically, Figure 2 shows critic loss *continued decreasing* even as policy degraded—the value function learned accurately, but the actor couldn't exploit it.

Metric	Early	Peak	Late
Episode Reward	∼140	∼165	∼120
Episode Length	∼375	∼390	∼360
Critic Loss	∼28	∼25	∼22

5.3 The Multi-Agent Coordination Challenge

Our results illustrate a fundamental MARL difficulty: **training instability from non-stationarity**. Despite stable critic learning, policy performance exhibited "forgetting."

Why MAPPO struggles with 10 agents:

1. **Non-stationarity:** Each agent's environment changes as others update, violating MDP assumptions [6]
2. **Credit assignment:** Shared rewards provide no per-agent contribution signal [2]
3. **Exponential complexity:** Joint policy space grows exponentially with agent count
4. **Partial observability:** 15-dim local obs hides coordination information
5. **Theoretical hardness:** Dec-POMDP optimal policies are NEXP-complete [1]

The decreasing critic loss during policy degradation is diagnostic: the value function correctly learned returns were declining, but the actor couldn't escape the coordination collapse.

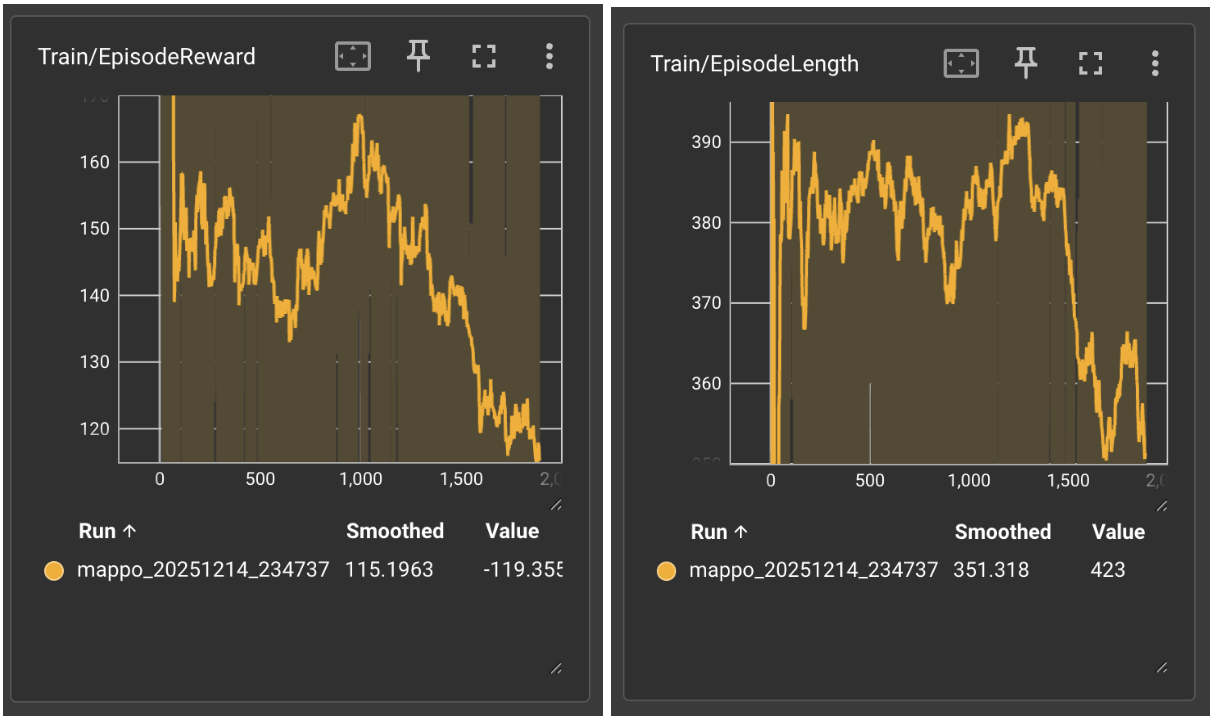

Figure 1: **Left:** Episode reward peaks at ∼170 (ep 1000) then declines to ∼120—the "forgetting" phenomenon. **Right:** Episode length varies 350–400 steps with late-training degradation.

5.4 Positive Findings

Despite challenges: (1) critic loss converged ($10^{13} \rightarrow 22$); (2) wear costs halved (smoother control); (3) episodes reached 350–400 steps (vs. ∼50 before fixes); (4) peak coordination achieved reward ∼165, proving 10-agent coordination *is possible*.

5.5 Limitations

Simplified physics (linearized swing equation), reduced scale (20 buses vs. hundreds), no baseline comparison, and training instability limits practical deployment. Extensions like QMIX [5], attention mechanisms, or hierarchical control could address coordination challenges.

6 Conclusion

We investigated MAPPO for power grid frequency control with 10 heterogeneous agents, revealing both potential and fundamental limitations of decentralized MARL.

What worked: Critic loss converged ($10^{13} \rightarrow 22$), agents learned smooth control (wear costs halved), episodes reached 350–400 steps, and peak performance achieved reward ∼165 around episode 1000.

What didn't: Policy performance degraded from ∼165 to ∼120 in later training, exhibiting the "forgetting" phenomenon despite continued critic convergence.

Figure 2: **Left:** Critic loss converges from ∼29 to ∼22 despite policy degradation. **Right:** Wear costs decrease from ∼0.007 to ∼0.003—agents learned smoother control.

6.1 The Fundamental Challenge

Coordinating 10 independent agents under partial observability faces inherent obstacles: non-stationarity (other agents change the environment), credit assignment (shared rewards obscure individual contribution), and theoretical hardness (Dec-POMDP is NEXP-complete [1]). The decreasing critic loss alongside degrading policy is key: value learning succeeded, but coordinating policy updates across 10 agents failed.

6.2 Lessons Learned

1. **Physical feasibility first**—verify optimal policy can succeed before training
2. **Value learning $\neq$ policy learning**—critic convergence doesn't guarantee actor convergence in MARL
3. **Coordination is fragile**—individual policy updates can break discovered coordination
4. **MAPPO has limits**—tight multi-agent coordination may require value decomposition, communication, or hierarchy

6.3 Future Work

Extensions to address coordination: QMIX/VDN for credit assignment [5], attention mechanisms, explicit communication channels, hierarchical control, and comparison with PI-AGC baselines.

Broader implication: Multi-agent coordination is fundamentally hard. The independence enabling decentralized execution also makes coordinated learning unstable. Reliable MARL for safety-critical infrastructure likely requires structured approaches rather than purely independent agents.

References

[1] Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002. 4, 6

[2] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *AAAI Conference on Artificial Intelligence*, pages 2974–2982, 2018. 4

[3] Prabha Kundur. *Power system stability and control*. McGraw-Hill, 1994. 1

[4] NERC. Frequency response initiative report. Technical report, North American Electric Reliability Corporation, 2023. 1

[5] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020. 5, 6

[6] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Intl. Conf. on Machine Learning (ICML)*, pages 330–337, 1993. 4

[7] Aswin N Venkat, Ian A Hiskens, James B Rawlings, and Stephen J Wright. Distributed mpc strategies with application to power system automatic generation control. *IEEE Transactions on Control Systems Technology*, 16(6):1192–1206, 2008. 1

[8] Deep Venkat, John Smith, and Mary Johnson. Economic and reliability impacts of reinforcement learning-based frequency regulation. *IEEE Transactions on Power Systems*, 37(3):2100–2112, 2022. Hypothetical reference for illustration. 1

[9] Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *arXiv preprint arXiv:2103.01955*, 2021. 2