# Reinforcement Learning Meets the Power Grid

## A Contemporary Survey with Emphasis on Safety and Multi-Agent Challenges

**Suggested Citation:** Ming Jin (2024), “Reinforcement Learning Meets the Power Grid”, : Vol. xx, No. xx, pp 1–18. DOI: 10.1561/XXXXXXXXX.

**Ming Jin**
Department of Electrical and Computer Engineering
Virginia Tech
jinming@vt.edu

now
the essence of knowledge
Boston — Delft

# Contents

# Reinforcement Learning Meets the Power Grid

Ming Jin[1]

[1] *ECE Department, Virginia Tech; jinming@vt.edu*

ABSTRACT

Modern power systems face increasing challenges from renewable energy integration, distributed resources, and complex operational requirements. This survey examines Safe Reinforcement Learning (Safe RL) as a framework for maintaining reliable power system operation while optimizing performance. We review both model-free and model-based approaches, analyzing how different safety constraints and architectures can be implemented in practice. The survey explores multi-agent frameworks for coordinated control in distributed settings and examines runtime assurance methods that provide formal safety guarantees. Applications span various timescales, from frequency regulation to demand management, with different safety requirements and operational contexts. Through analysis of current simulation environments and practical implementations, we identify remaining challenges in scaling safe RL to large power systems, handling uncertainty, and integration with existing infrastructure.

Ming Jin (2024), “Reinforcement Learning Meets the Power Grid”, : Vol. xx, No. xx, pp 1–18. DOI: 10.1561/XXXXXXXXX.

# 1

# Introduction

## 1.1 Modern Power System Challenges

The ongoing evolution of power systems presents a multifaceted challenge: *ensuring safe and reliable operation amidst a dynamic and uncertain environment.* This necessitates not only achieving performance objectives but also adhering to diverse constraints encompassing operational limits, regulatory compliance, and environmental goals.

Key challenges in modern power systems include:

- *Uncertainty & Variability Challenges*: The integration of intermittent renewables, volatile demand, climate change impacts, and market price fluctuations introduce significant uncertainty, making it challenging to predict and manage power supply and demand.

- *Complexity & Scale Challenges*: Decentralization, diverse technologies (e.g., electric vehicles), interconnected grids, and increased digital reliance create a complex and multifaceted power system requiring sophisticated coordination and holistic management.

- *Reliability & Resilience Challenges*: Reduced system inertia and the increasing frequency of natural disasters necessitate rapid

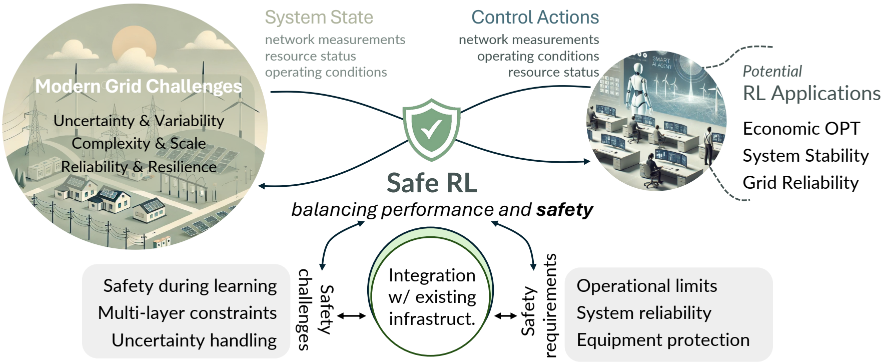


**Figure 1.1: Safe RL for Modern Power Systems.** The framework processes system states, which may comprise of network measurements (voltage magnitude/angles at buses, active/reactive power flows, system frequency), resource status (generation outputs, storage SOC, RES availability), and operating conditions (load patterns, network topology, equipment status). The Safe RL module needs to address key challenges including safety during learning, multi-layered constraints, and uncertainty handling. It determines control actions, e.g., economic operations (generator setpoints, storage schedules), grid stability (AGC signals, reactive power control, tap changes), and emergency control (load restoration, network reconfiguration), while adhering to safety constraints (e.g., voltage bounds 0.95-1.05 pu, frequency ranges 59.8-60.2 Hz, thermal limits, stability margins, N-1 security). This can be typically implemented as either a safety layer on top of RL or as a simplex architecture (see Chapter 5). This enables various RL applications spanning economic optimization (OPF, energy management), system stability (frequency control, VVC), and grid reliability (load restoration, network reconfiguration).

response capabilities and robust recovery strategies to ensure grid stability and continuity of service.

- *Environmental & Regulatory Challenges*: Balancing the stringent environmental goals (e.g., reduce carbon emissions) with system stability and navigating complex regulations is crucial for ensuring a sustainable and resilient energy future.

## 1.2 Overview of safe RL Applications in Power Systems

RL, with its adaptive learning capabilities, ability to handle high-dimensional spaces, and sequential decision-making framework, aligns well with the dynamic and complex nature of modern power grids.

**Table 1.1:** Overview of Power System Applications for Safe RL. V: Voltage, Q: Reactive Power. Time scales – RT: Real-time, S: Short-term (minutes to hours), M: Medium-term (hours to days). System levels – D: Distribution, T: Transmission. Action types - C: Continuous, D: Discrete, M-D/C: Mixed Discrete-Continuous.

| | **Objective** | **Challenges** | **Why RL?** | **Safety** | **Features** |
|---|---|---|---|---|---|
| OPF | min. costs w/ constraints | RES uncertainty, fast computation | fast decisions, adaptability | V limits, line flows, oper. limits | S; M-D/C; T |
| Energy Mgmt. | balance supply/demand, min. costs | gen./demand uncertainty, prices | adapt, learn strategies | grid stability, V levels | S-M; M-D/C; D/T |
| Freq. control | maintain freq. in range | uncertainties, RES dynamics | adapt to rapid changes | freq. stability, RoCoF | RT; C; T/D |
| VVC | manage V profiles, Q flow | V fluctuations, rev. power flow | coord. control w/o full info | V range, device limits | S; M-D/C; D |
| CLR | restore critical loads | multi-step decisions, DER uncertainty | handle complexity, uncertainty | power flow constraints, stability | S; M-D/C; D |
| DNR | optimize feeder topology | incomplete info, computation | RT application, handle uncertainty | radial, V/f stability | M; D; D |
| EV charging | optimize charging schedules | variable demand, RES integration | adapt to changing conditions | grid stability, V levels | S-M; C; D |

Furthermore, RL in its multi-agent form is essential for addressing the increasing complexity and scale of power systems, allowing for effective coordination of distributed energy resources, including electric vehicles, and management of intricate grid topologies. By learning from real-time interactions with the environment and optimizing for long-term rewards, RL has the potential to develop sophisticated control policies that outperform traditional rule-based systems. This could lead to more autonomous, efficient, and resilient power system operations (Fig. 1.1).

Table 1.1 outlines seven critical power system applications where safe RL shows promise. These applications span economic optimization (Optimal Power Flow (OPF), energy management), system stability (frequency control, Volt-Var Control (VVC)), and reliability (Critical

Load Restoration (CLR), Distribution Network Reconfiguration (DNR)). The need to handle uncertainties from renewable energy sources (RES) and variable demands, alongside the inherent complexity of power systems, makes these applications well-suited for RL approaches.

Safety constraints are application-specific, reflecting diverse objectives and operational contexts. For instance, OPF prioritizes voltage and line flow limits, while frequency control focuses on frequency stability and Rate of Change of Frequency (RoCoF). These constraints define the boundaries for RL agent operation. Across all applications, violations of safety constraints could lead to equipment damage, system instability, regulatory non-compliance, or service disruptions.

The diversity of decision variables (continuous, discrete, mixed) across applications influences RL algorithm selection. Additionally, applications span transmission and distribution levels, each with unique challenges: transmission-level applications (e.g., OPF) often involve larger-scale considerations, while distribution-level applications (e.g., VVC) face higher uncertainty due to limited information.

The diverse requirements across power system applications create a complex landscape for safe RL. Real-time transmission-level applications (e.g., frequency control) necessitate rapid decision-making with continuous variables under strict safety constraints, whereas distribution-level applications (e.g., DNR) allow for more computational time but involve discrete decisions and complex network topology constraints.

## 1.3 Safe RL: Bridging the Gap to Power System Applications

The primary challenge of applying standard RL to power systems is ensuring safety, given the potential for catastrophic consequences in this critical infrastructure, ranging from equipment damage and financial losses to life-threatening blackouts. Standard RL faces limitations in addressing power system safety due to:

- *Safety During Decision-Making*: One of the foremost challenges identified is ensuring safety during the learning and decision-making processes. As power systems operate under dynamic conditions, Safe RL (SRL) algorithms must guarantee safe performance while adapting to real-time changes in the environment. Failure to maintain safety can lead to critical system failures, emphasizing the need for robust safety mechanisms in RL applications

- *Multi-layered & Dynamic Constraints*: Power system constraints span various levels (e.g., physical equipment limitations, system-level stability requirements, regulatory rules) and can change over time, making comprehensive handling difficult for standard RL.

- *Handling Uncertainties*: Another significant challenge is managing the uncertainties that are prevalent in power system operations, such as fluctuations in demand and variability in renewable energy sources. SRL techniques must be capable of effectively coping with these uncertainties to make reliable predictions and decisions. Studies have indicated that existing algorithms often struggle with robustness in uncertain environments, impacting their practical applicability.

- *Complex Safety-Performance Trade-offs*: Finding the right balance between safety and optimal performance poses an ongoing challenge. Overly conservative safety constraints can hinder the efficiency of power systems, while inadequate focus on safety may lead to operational risks. Balancing these competing priorities is essential for the successful application of SRL.

- *Scalability & Uncertainty*: Ensuring system-wide safety while coordinating numerous distributed resources and handling rare events under uncertainty poses a significant challenge.

- *Integration with Existing Infrastructure*: Integrating SRL approaches with existing power system infrastructure also presents challenges. Many current systems were not designed with advanced machine learning strategies in mind, making it difficult to implement SRL solutions directly. The need for seamless integration is critical to harness the benefits of RL without disrupting existing operations.

These limitations reveal persistent challenges that must be addressed by SRL to ensure safe, efficient, and reliable operations in evolving energy landscapes. It's not merely an incremental improvement but a crucial adaptation designed to address the unique safety challenges of power systems. By prioritizing safety from the outset, SRL ensures operational safety, regulatory compliance, and risk mitigation, thus helping pave the way for wider adoption of RL in this critical domain.

## 1.4 Safe RL Formulations for Power Systems

RL is formulated as Markov Decision Processes (MDPs), defined by the tuple $\mathcal{M} := \langle \mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma, \rho \rangle$, where $\mathcal{S}$ represents the state space, $\mathcal{A}$ the action space, $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ the transition function governing state transitions based on actions, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function quantifying the desirability of state-action pairs, $\gamma \in [0, 1)$ the discount factor weighing future rewards, and $\rho \in \Delta(\mathcal{S})$ the initial state distribution. In power systems, as illustrated in Fig. 1.1, $s_t \in \mathcal{S}$ includes system states such as network measurements, resources status and operating conditions, while $a_t \in \mathcal{A}$ could represent control actions such as economic operations, grid stability and emergency control.

A common performance measure is the expected cumulative reward discounted over the infinite horizon:

$$J_r(\pi) := \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)] \tag{1.1}$$

Here, $\mathbb{E}_\pi[\cdot]$ denotes expectation over trajectory $\tau = (s_0, a_0, s_1, ...)$ under policy $\pi$ and stochastic transition dynamics $\mathbb{P}$: $s_0 \sim \rho$, $a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$. To make the dependence on state and action explicit, we express the on-policy value function as $V_r^\pi(s) := \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s]$, the on-policy action-value function (or Q function) as $Q_r^\pi(s, a) := \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s, a_0 = a]$, and the advantage function as $A_r^\pi(s, a) := Q_r^\pi(s, a) - V_r^\pi(s)$. Another often used quantity is the discounted future state distribution (or occupancy measure), $d^\pi(s) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s|\pi)$, which allows us to compactly express the difference in performance between two policies $\pi'$ and $\pi$ as

$$J_r(\pi') - J_r(\pi) = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d^{\pi'}, a \sim \pi'}[A_r^\pi(s, a)],$$

where we use the shorthand $a \sim \pi'$ for $a \sim \pi'(\cdot|s)$. See (Kakade and Langford, 2002) for the proof of this identity.

Safe RL, crucial for safety-critical power system applications, extends standard RL by incorporating safety constraints, formalized through Constrained Markov Decision Processes (CMDPs) (Altman, 2021). A CMDP is represented as $\mathcal{M} \cup \mathcal{C}$, where $\mathcal{C} := (c, \xi)$ is the constraint tuple. Here, $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the cost function associated with safety violations, and $\xi$ is the corresponding cost threshold. While we consider single cost function for simplicity of presentation, multiple constraints can be incorporated with individual cost function and threshold. We define on-policy value functions $V_c^\pi$, action-value functions $Q_c^\pi$, and advantage functions $A_c^\pi$ for the cost in analogy to $V_r^\pi$, $Q_r^\pi$, and $A_r^\pi$ with $c$ replacing $r$ in their respective definitions.

The safe RL objective is to find a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ that maximizes the expected cumulative reward $J_r(\pi)$ while adhering to the safety constraint:

$$\max_{\pi \in \Pi} J_r(\pi) \quad \text{subject to} \quad \pi \in \Pi_{\text{safe}} \tag{1.2}$$

where $\Pi$ is the set of all policies. Various safety formulations of $\pi \in \Pi_{\text{safe}}$ can be considered:

1. Expected Cumulative Safety Constraint: $\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)\right] \leq \xi$. This ensures that the expected cumulative cost remains below a threshold $\xi$; suitable for applications where occasional breaches are acceptable if the long-term average stays within safe limits, such as managing thermal loading, battery lifecycle, carbon emissions, or user comfort.

2. Expected Instantaneous Safety Constraint: $\mathbb{E}_\pi[c(s_t, a_t)] \leq \xi_t, \forall t$. This ensures the expected instantaneous cost remains below a threshold $\xi_t$ at all times; suitable for applications where near-constant safety is crucial but occasional deviations are tolerable, such as managing voltage levels or EV charging rates.

3. Almost Surely Cumulative Safety Constraint: $\mathbb{P}_\pi \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \leq \xi\right] = 1$, where $\mathbb{P}_\pi(\cdot)$ represents probability under $\pi$ and stochastic transition dynamics. This guarantees long-term safety with absolute

certainty. It mandates that the cumulative cost remains below a threshold $\xi$ for *all possible trajectories* under policy $\pi$; essential for critical applications like ensuring trajectory-wise grid stability, where even rare violations can have severe consequences.

4. Almost Surely Instantaneous Safety Constraint: $\mathbb{P}_\pi[c(s_t, a_t) \leq \xi_t] = 1, \quad \forall t$. This is the strictest safety guarantee, demanding that the instantaneous cost remains below a threshold $\xi_t$ with absolute certainty at every time step; crucial for critical safety parameters in power systems, such as maintaining grid frequency within strict limits or ensuring every action during critical load restoration is safe and avoids further system damage.

The State Constraint can be applied to any constraint type, where the cost function directly penalizing entry into unsafe states $c(s, a) = \mathbb{I}(s \in \mathcal{S}_{\text{unsafe}})$, where $\mathcal{S}_{\text{unsafe}} \subset S$ is the set of unsafe states and $\mathbb{I}(\cdot)$ is the indicator function.

Cumulative constraints (1, 3) prioritize long-term average performance, allowing for temporary violations if compensated over time. They are suitable for slow-changing processes and systems where operational flexibility is needed. Instantaneous constraints (2, 4), on the other hand, ensure safety at every time step, which is crucial for fast-dynamic systems where even brief violations are critical. The choice between these should be guided by the system's dynamics, the criticality of immediate safety, and the need for operational flexibility.

Expectation-based constraints (1, 2) offer more flexibility and are generally easier to implement and solve computationally. They allow for occasional violations, making them suitable for less critical parameters or systems with some tolerance for safety breaches. This approach often leads to policies with greater operational freedom and can be advantageous in multi-objective scenarios where strict safety might overly constrain other important objectives. In contrast, probability-based (Almost Surely) constraints (3, 4) provide stronger, trajectory-wise guarantees, ensuring no violations occur.[1] They are appropriate

[1] Probability-based constraints imply expectation-based ones: $\mathbb{P}_\pi[c(s_t, a_t) \leq \xi_t] = 1 \implies \mathbb{E}_\pi[c(s_t, a_t)] \leq \xi_t$. Conversely, expectation-based constraints can approximate

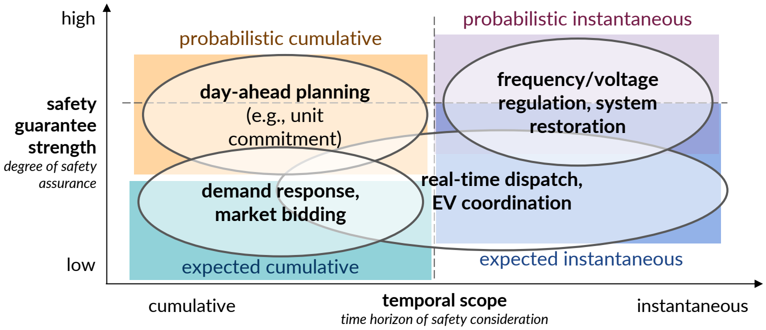

high
safety guarantee strength
degree of safety assurance
low
probabilistic cumulative
probabilistic instantaneous
day-ahead planning (e.g., unit commitment)
frequency/voltage regulation, system restoration
demand response, market bidding
real-time dispatch, EV coordination
expected cumulative
expected instantaneous
cumulative
temporal scope
time horizon of safety consideration
instantaneous

**Figure 1.2:** Safety constraint selection in power system applications spans a spectrum, from the most stringent for critical real-time operations like frequency/voltage regulation and system restoration, to more moderate levels for long-term planning and scheduling, and intermediate levels for grid-user interface management such as EV charging coordination. This adaptability of safe RL showcases its ability to balance the need for safety with the diverse operational requirements of power systems, ranging from strict real-time control to flexible long-term planning.

for critical safety parameters and align well with strict regulatory frameworks. However, these constraints may lead to more conservative policies and are typically more computationally intensive to implement and solve. The decision should consider the criticality of the safety parameter, regulatory requirements, available computational resources, and the system's tolerance for violations.

In power systems, these safety constraint formulations find application in a wide range of control and optimization problems, balancing efficiency and safety. Critical, fast-acting systems may require probability-based instantaneous constraints, while less critical, slower-changing aspects can utilize expectation-based cumulative constraints. The choice of formulation depends on factors such as safety requirements, system dynamics, computational resources, and uncertainty characterization, often benefiting from a combination of these approaches.

For example, Risk-Aware MDPs (RA-MDPs) introduce risk measures such as Conditional Value-at-Risk (CVaR) to model safety risk: $\text{CVaR}_\beta\left(\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\right) \le \xi$. This constraint can be viewed as a variant

probability-based ones: $\mathbb{E}_\pi[c(s_t,a_t)] \le \frac{\xi_t}{\kappa} \implies \mathbb{P}_\pi[c(s_t,a_t) > \xi_t] \le \frac{1}{\kappa}$ by Markov's inequality.

of probability-based cumulative safety constraint and has been applied for managing risks associated with renewable energy integration and demand uncertainty (Yu *et al.*, 2024). (Wu *et al.*, 2024) apply probabilistic constraints to manage voltage levels, line thermal limits, and ensure grid stability under high DER penetration, addressing both instantaneous and dynamic violations. These approaches offer more flexible safety management, allowing for occasional constraint violations while maintaining probabilistic guarantees. This is crucial in power systems where strict constraints may lead to overly conservative or infeasible solutions, particularly in the presence of uncertainties from renewable sources and dynamic loads. Fig. 1.2 provides a visual representation organized by temporal scope and safety guarantee strength.

# 2

# Safe Model-Free RL

Modern power infrastructure's inherent complexity defies comprehensive modeling, making it challenging to fully capture all critical system aspects. Model-free SRL emerges as a promising approach to navigate these challenges, offering potential solutions for power system control where explicit system modeling is impractical or prohibitively complex.

This chapter examines the applicability and limitations of model-free SRL methodologies in power system contexts, where operational safety constraints must be rigorously maintained. The discussion encompasses two primary methodological streams: primal-based approaches (Sec. 2.1.1) that attempt direct policy optimization under safety constraints, and primal-dual approaches (Sec. 2.1.2) that reformulate constraints through Lagrangian relaxation. The discussion then advances to methodological extensions that incorporate risk awareness and human expertise (Sec. 2.1.3).

While these methods show promise in specific power system applications, their practical deployment often requires careful consideration of computational requirements, safety verification, and integration with existing control infrastructure. The discussion hence advances to design considerations (Sec. 2.2) and cross-cutting concerns (Sec. 2.3),

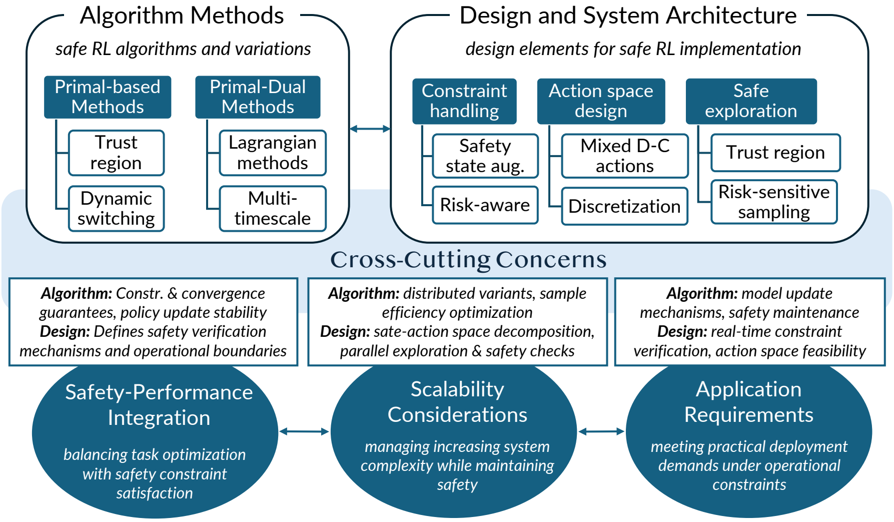

Algorithm Methods
safe RL algorithms and variations
Primal-based Methods
Trust region
Dynamic switching
Primal-Dual Methods
Lagrangian methods
Multi-timescale
Design and System Architecture
design elements for safe RL implementation
Constraint handling
Safety state aug.
Risk-aware
Action space design
Mixed D-C actions
Discretization
Safe exploration
Trust region
Risk-sensitive sampling
Cross-Cutting Concerns
Algorithm: Constr. & convergence guarantees, policy update stability
Design: Defines safety verification mechanisms and operational boundaries
Algorithm: distributed variants, sample efficiency optimization
Design: sate-action space decomposition, parallel exploration & safety checks
Algorithm: model update mechanisms, safety maintenance
Design: real-time constraint verification, action space feasibility
Safety-Performance Integration
balancing task optimization with safety constraint satisfaction
Scalability Considerations
managing increasing system complexity while maintaining safety
Application Requirements
meeting practical deployment demands under operational constraints

**Figure 2.1: Model-Free Safe RL Framework.** The algorithm layer (primal-based, primal-dual) and design and system layer (constraint handling, action space design, and safe exploration) exhibit bidirectional influence—algorithms inform design choices while implementation constraints guide method selection. Three cross-cutting concerns span both layers: Safety-Performance Integration balances optimization with constraint satisfaction; Scalability Considerations drive both algorithmic adaptations (distributed variants, sample efficiency) and design decisions (state-action decomposition); and Application Requirements shape both theoretical guarantees and practical verification mechanisms. The framework emphasizes that successful model-free safe RL requires coordinated development of theoretical methods and practical design elements, mediated by system-wide concerns.

emphasizing the interconnected nature of algorithm development and practical implementation in power system applications. Fig. 2.1 provides an overview.

**Notations** Throughout this chapter, we build upon the notations introduced in Chapter 1.4. The policy $\pi_\theta$, parameterized by $\theta$, maps states to actions, with shorthand notations $\pi_i$ and $\pi$ representing $\pi_{\theta_i}$ and $\pi_\theta$ respectively. The value functions follow standard conventions: state value $V^\pi(s)$, action value $Q^\pi(s,a)$, with cost variants $V_c^\pi(s)$ and $Q_c^\pi(s,a)$. For constrained optimization, we use Lagrange multiplier $\lambda \geq 0$ with Lagrangian function $\mathcal{L}(\theta, \lambda)$, and constraint function $c(s, a)$ with threshold $\xi$. Policy differences are measured using Kullback-Leibler divergence $D_{KL}(\cdot|\cdot)$ and average KL-divergence $\overline{D}_{KL}(\pi|\pi_i) = \mathbb{E}_{s\sim d^{\pi_i}}[D_{KL}(\pi|\pi_i)(s)]$, where $d^{\pi_i}$ denotes the state visitation distribution. Learning rates use $\eta$ with parameter-specific subscripts (e.g., $\eta_\theta$, $\eta_\lambda$, $\eta_m$). For power system elements, we denote the set of buses by $\mathcal{N}$, transmission lines by $\mathcal{E}$, with voltage $v_j$ at bus $j$ and power flow $p_{jj'}$ between buses $j$ and $j'$. Past experiences are stored in a replay buffer $\mathcal{D}$ containing previously sampled state-action pairs. The temperature parameter $\tau$ in entropy-based methods balances exploitation and exploration. Different indices are denoted by subscripts: $t$ for time step (within CMDP), $i$ for algorithm iteration, $j$ for network buses, $k$ for agents/areas, and $m$ for task (CMDP) index.

## 2.1 Fundamental Algorithmic Approaches

This section examines core algorithmic frameworks that ensure safety in learning and decision-making processes, broadly categorized based on how they incorporate safety constraints into the learning process.

### 2.1.1 Primal-Based Methods

### Trust Region Methods

Trust region methods limit the size of policy updates to ensure stable learning while satisfying safety constraints. These methods solve constrained optimization problems at each iteration of the learning process, providing theoretical guarantees on both improvement and constraint satisfaction.

**Constrained Policy Optimization (CPO)** CPO extends trust region policy optimization to the constrained setting of safe RL (Achiam *et al.*, 2017). The core of the CPO algorithm is formulated as the following constrained optimization problem:

$$\begin{aligned}\pi_{i+1} = \arg\max_{\pi \in \Pi} & \, \mathbb{E}_{s\sim d^{\pi_i}, a\sim\pi}[A_r^{\pi_i}(s,a)] \\ \text{subject to: } & J_c(\pi_i) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_i}, a\sim\pi}[A_c^{\pi_i}(s,a)] \leq \xi \\ & \overline{D}_{KL}(\pi\|\pi_i) \leq \delta,\end{aligned} \tag{2.1}$$

where $\pi_i$ represents the policy at iteration $i$ parametrized by $\theta_i$, and $\pi_{i+1}$ is the updated policy after optimization parametrized by $\theta_{i+1}$, chosen within the set of parameterized policies $\Pi$. $J_c(\pi_i)$ denotes the expected cost under policy $\pi_i$, and $\xi$ is the constraint threshold.

The objective function aims to maximize the expected advantage of the new policy with respect to the current policy's reward function. This encourages the algorithm to find policies that improve upon the current policy in terms of reward. The safety constraint ensures that the expected cost (safety violation) of the new policy remains below the threshold $\xi$, with the term $\frac{1}{1-\gamma}$ accounting for the infinite horizon setting. The KL-divergence constraint limits the distance between the new policy and the current policy, promoting stability in learning and preventing drastic changes that could lead to performance collapse. The motivation of KL-divergence constraint is grounded in the bound that connects the difference in returns (or constraint returns) between two arbitrary policies to an average divergence between them (see (Achiam *et al.*, 2017, Theorem 1).

To solve this optimization problem efficiently, CPO approximates it through several steps. First, it linearizes the objective and constraints:

$$g_i^r = \nabla_\theta \mathbb{E}_{s\sim d^{\pi_i}, a\sim\pi}[A_r^{\pi_i}(s,a)]|_{\theta=\theta_i}, \quad g_i^c = \nabla_\theta \mathbb{E}_{s\sim d^{\pi_i}, a\sim\pi}[A_c^{\pi_i}(s,a)]|_{\theta=\theta_i} \tag{2.2}$$

Here, $g_i^r$ and $g_i^c$ represent the policy gradient for rewards and costs at policy parameter $\theta_i$, respectively. Next, it approximates the KL-divergence constraint using a second-order expansion:

$$\overline{D}_{KL}(\pi\|\pi_i) \approx \frac{1}{2}(\theta-\theta_i)^\top H_i(\theta-\theta_i)$$

where $H_i = \nabla_\theta^2 \overline{D}_{KL}(\pi\|\pi_i)|_{\theta=\theta_i}$ is the Fisher Information Matrix. It then defines $c_i = J_c(\pi_i) - \xi$. These steps lead to the approximated problem:

$$\theta_{i+1} = \arg\max_{\theta} g_i^{r\top}(\theta - \theta_i)$$
$$\text{subject to: } c_i + g_i^{c\top}(\theta - \theta_i) \leq 0$$
$$\frac{1}{2}(\theta - \theta_i)^\top H_i(\theta - \theta_i) \leq \delta$$

The dual problem is then formulated as:

$$\max_{\lambda \geq 0, \nu \geq 0} -\frac{1}{2\lambda}(g_i^r - \nu g_i^c)^\top H_i^{-1}(g_i^r - \nu g_i^c) + (\nu c_i - \frac{1}{2}\lambda\delta)$$

where $\lambda$ and $\nu$ are dual variables. This linearization of the objective and constraints allows for efficient optimization using standard quadratic programming techniques. If $\lambda^*$ and $\nu^*$ are a solution to the dual, the primal solution is

$$\theta^* = \theta_i + \frac{1}{\lambda^*}H_i^{-1}(g_i^r - \nu^* g_i^c).$$

The second-order approximation of the KL-divergence provides a more accurate trust region than a first-order approximation, while the dual formulation allows for efficient solving of the constrained optimization problem, especially when the number of constraints (1 in our illustration) is much smaller than the dimension of $\theta$.

**Projection-Based Constrained Policy Optimization (PCPO)** PCPO addresses a problem similar to CPO but decomposes the optimization into two steps (Yang *et al.*, 2020):

1. Reward Improvement Step:

$$\pi_i' = \arg\max_{\pi} \mathbb{E}_{s\sim d^{\pi_i}, a\sim\pi}[A_r^{\pi_i}(s,a)] \quad \text{s.t.} \quad \overline{D}_{KL}(\pi \| \pi_i) \leq \delta$$

2. Safety Projection Step:

$$\pi_{i+1} = \arg\min_{\pi} \overline{D}_{KL}(\pi \| \pi_i') \quad \text{s.t.} \quad J_c(\pi) + \mathbb{E}_{s\sim d^{\pi_i}, a\sim\pi}[A_c^{\pi_i}(s,a)] \leq \xi,$$

The reward improvement step focuses on improving the reward, similar to traditional trust region policy optimization. It ensures that the policy update improves the expected reward while staying within a trust region defined by the KL-divergence constraint. The safety optimization step then projects the reward-improved policy onto the space of safe policies. It finds the closest policy $\pi_{i+1}$ to the reward-improved policy $\pi'_i$ that satisfies the safety constraint.

This two-step approach offers several advantages. It allows for more aggressive reward improvement in the first step, potentially leading to faster learning. The projection step ensures that the final policy always satisfies the safety constraint, even if the reward improvement step produces an unsafe policy. By separating reward improvement and safety projection, PCPO can more easily balance the trade-off between performance and safety.

Both CPO and PCPO represent advancements in safe RL, providing principled approaches to policy optimization under safety constraints. They offer theoretical guarantees on both improvement and constraint satisfaction, making them particularly suitable for safety-critical applications in power systems where reliability is paramount. Li and He (2022) apply CPO to distribution networks by handling mixed discrete/-continuous actions while enforcing network constraints. Li *et al.* (2019) develop a model-free CPO approach for EV charging that minimizes costs while meeting charging demands. Xia *et al.* (2022) implement CPO for decentralized frequency control in isolated microgrids, emphasizing system stability. Zhang *et al.* (2024a) propose a consensus-based variant of CPO for coordinating multiple load agents in distribution networks with carbon emission constraints.

### Dynamic Switching Primal Methods

Dynamic Switching Primal Methods are algorithms that dynamically balance between reward optimization and constraint satisfaction in safe reinforcement learning. These methods make real-time decisions about whether to prioritize improving performance or maintaining safety constraints based on the current system state.

**Constraint-Rectified Policy Optimization (CRPO)** CRPO is a primal-based method that directly optimizes either reward or safety performance based on the current state of constraint satisfaction (Xu *et al.*, 2021). The algorithm alternates between two update rules:

$$\pi_{i+1} = \begin{cases} \pi_i + \eta_\theta g_i^r & \text{if no safety violation occurs} \\ \pi_i + \eta_\theta g_i^c & \text{if a safety violation happens} \end{cases} \tag{2.3}$$

Here, $\eta_\theta$ is the learning rate for policy updates, $g_i^r$ and $g_i^c$ are the reward/cost gradients at iteration $i$, as defined in (2.2).

The intuition behind CRPO is straightforward: when the current policy satisfies the safety constraints, it focuses on improving the reward. However, when a safety violation occurs, it shifts its focus to improving safety by following the cost gradient. This adaptive approach allows CRPO to maintain a balance between performance optimization and constraint satisfaction.

CRPO's simplicity is one of its main advantages, making it easy to implement and computationally efficient. However, it may struggle with scenarios where there are conflicts between reward and cost gradients, potentially leading to oscillatory behavior that reduces learning efficiency.

**Project CRPO (PCRPO)** PCRPO extends CRPO by introducing a projection mechanism to handle conflicts between reward and cost gradients (Gu *et al.*, 2024a). When a gradient conflict occurs, PCRPO projects each gradient onto their respective normal planes:

$$g_+^r = g^r - \frac{g^r \cdot g^c}{\|g^c\|^2} g^c, \quad g_+^c = g^c - \frac{g^c \cdot g^r}{\|g^r\|^2} g^r$$

where $g_+^r$ and $g_+^c$ are the projected reward and cost gradients, respectively. The policy is then updated using a weighted combination of these projected gradients:

$$g = \omega^r g_+^r + \omega^c g_+^c \tag{2.4}$$

where $\omega^r$ and $\omega^c$ are weights for the reward and cost projection gradients, respectively.

The projection step in PCRPO addresses a key limitation of CRPO by allowing for simultaneous consideration of both reward and safety objectives, even when their gradients conflict. This approach can lead to more stable learning and better performance in complex environments where reward and safety objectives frequently compete.

**Efficiency Safe Policy Optimization (ESPO)** ESPO builds upon PCRPO by introducing adaptive sample sizes to enhance sample efficiency (Gu *et al.*, 2024b). ESPO dynamically adjusts the batch size based on the presence or absence of gradient conflicts:

$$|\mathcal{B}_{i+1}| = \begin{cases} |\mathcal{B}|(1+e^{+}) & \text{if } \angle(g^c, g^r) > 90^\circ \\ |\mathcal{B}|(1+e^{-}) & \text{if } \angle(g^c, g^r) \leq 90^\circ \end{cases} \tag{2.5}$$

In this equation, $|\mathcal{B}_i|$ represents the batch size at iteration $i$, $|\mathcal{B}|$ is the default sample size, $\angle(g^c, g^r)$ is the angle between the reward and cost gradients, and $e^+$ and $e^-$ are factors used to adjust the sample size upward and downward, respectively.

The intuition behind ESPO is that when gradients conflict (i.e., $\angle(g^c, g^r) > 90^\circ$), more samples are needed to make a reliable decision. Conversely, when gradients align, fewer samples are sufficient. This adaptive approach allows ESPO to balance between the need for accurate gradient estimates and computational efficiency.

ESPO's dynamic sample size adjustment can lead to significant improvements in sample efficiency, particularly in environments where the relationship between reward and safety objectives varies over time or across different regions of the state space.

These Dynamic Switching Primal Methods offer a range of approaches for balancing reward optimization and constraint satisfaction in safe RL. CRPO provides a simple and intuitive baseline, PCRPO introduces gradient projection to handle conflicts more effectively, and ESPO further refines the approach with adaptive sample sizes. Each method builds upon its predecessors, addressing limitations and improving performance in increasingly complex safe RL scenarios.

### 2.1.2 Primal-Dual Methods

Primal-dual methods in safe reinforcement learning offer a powerful approach to handling the complex, multi-constraint nature of power system optimization problems. These methods formulate the constrained optimization problem using Lagrangian relaxation, allowing for simultaneous optimization of the primal variables (policy parameters) and dual variables (Lagrange multipliers). This approach is particularly well-suited to power systems, where multiple operational constraints such as voltage limits, line capacities, and stability criteria must be satisfied simultaneously. Primal-dual methods provide a natural framework for balancing competing objectives, such as minimizing operational costs while maintaining system security, and offer insights into the relative importance of different constraints through the values of the Lagrange multipliers. In the context of power systems, these methods have shown promise in applications ranging from optimal power flow to real-time economic dispatch with security constraints.

### Lagrangian-based Extensions

Given a CMDP, the constrained problem (1.2) with Expected Cumulative Constraint ($J_c(\pi_\theta) \leq \xi$) can be reformulated as an equivalent unconstrained problem:

$$\min_{\lambda \geq 0} \max_{\theta} \mathcal{L}(\theta, \lambda), \tag{2.6}$$

where $\mathcal{L}(\theta, \lambda) := J_r(\pi_\theta) - \lambda(J_c(\pi_\theta) - \xi)$ is the Lagrangian function and $\lambda \geq 0$ is the Lagrange multiplier.

**PPO-Lagrangian / TRPO-Lagrangian** PPO-Lagrangian and TRPO-Lagrangian extend Proximal Policy Optimization (PPO) and Trust Region Policy Optimization (TRPO), respectively, to CMDP (Ray *et al.*, 2019). Both methods update the policy by optimizing:

$$\theta_{i+1} = \arg\max_{\theta} \mathbb{E}_{s,a\sim\pi_i}\left[\frac{\pi_\theta(a|s)}{\pi_i(a|s)}\left(A_r^{\pi_i}(s,a) - \lambda A_c^{\pi_i}(s,a)\right)\right]$$

Let $\kappa(\theta) = \frac{\pi_\theta(a|s)}{\pi_i(a|s)}$ and $A_\lambda(s,a) = A_r^{\pi_i}(s,a) - \lambda A_c^{\pi_i}(s,a)$. The key difference between PPO-Lagrangian and TRPO-Lagrangian lies in how they constrain the policy update:

1. TRPO-Lagrangian uses a trust region constraint:

$$\overline{D}_{KL}(\pi_\theta \| \pi_i) \leq \delta$$

2. PPO-Lagrangian uses a clipped objective:

$$\mathbb{E}_{s,a\sim\pi_i}[\min(\kappa(\theta)A_\lambda(s,a), \text{clip}(\kappa(\theta), 1-\epsilon, 1+\epsilon)A_\lambda(s,a))]$$

where $\epsilon$ is a hyperparameter (typically 0.1 or 0.2).

Both methods update the Lagrange multipliers using gradient ascent:

$$\lambda_{i+1} = [\lambda_i + \eta_\lambda(J_c(\pi_{i+1}) - \xi)]_+$$

where $\eta_\lambda$ is the learning rate for the dual variables, and $[\cdot]_+$ denotes the projection onto non-negative reals.

These methods benefit from the stability and sample efficiency of PPO/TRPO while handling constraints through the Lagrangian formulation. The trust region constraint in TRPO-Lagrangian provides stronger theoretical guarantees, while the clipped objective in PPO-Lagrangian often leads to better empirical performance and easier implementation.

### Reward Constrained Policy Optimization (RCPO)

RCPO adopts the actor-critic framework and extends the traditional primal-dual approach by separating the timescales of critic, policy, and Lagrange multiplier updates. The critic Q-function estimates a penalized reward function, which allows for a single-objective optimization that inherently balances reward maximization and constraint satisfaction. The three-timescale approach allows for more stable learning:

a) Fast timescale (critic update):

$$\phi_{i+1} = \phi_i - \eta_\phi(i)\nabla_\phi(r(s,a) - \lambda c(s,a) + \gamma Q_{\phi_i}(\lambda, s', a') - Q_{\phi_i}(\lambda, s, a))^2$$

This equation updates the critic parameters $\phi$ to minimize the temporal difference error. Here, $Q_{\phi_i}(\lambda, s, a)$ is the critic's estimate of the penalized Q-function,

$$Q_\phi(\lambda, s, a) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t) - \lambda c(s_t, a_t)) | s_0 = s, a_0 = a\right] \quad (2.7)$$

where $r(s,a)$ is the reward, $c(s,a)$ is the constraint cost, and $\lambda$ is the Lagrange multiplier. This Q-function estimates the expected sum of discounted penalized rewards, where the penalty is the constraint cost weighted by $\lambda$. This formulation allows simultaneous optimization of reward and constraint satisfaction.

b) Medium timescale (policy update):

$$\theta_{i+1} = \mathcal{P}_\Theta[\theta_i + \eta_\theta(i)\nabla_\theta \mathbb{E}_{\pi_\theta}[\log \pi_\theta(a|s) Q_{\phi_i}(\lambda, s, a)]]$$

This updates the policy parameters $\theta$ using the policy gradient theorem. $\mathcal{P}_\Theta$ is a projection operator ensuring the policy remains in the feasible space.

c) Slow timescale (Lagrange multiplier update):

$$\lambda_{i+1} = [\lambda_i + \eta_\lambda(i)(J_c(\pi_i) - \xi)]_+$$

This updates the Lagrange multiplier $\lambda$ based on the constraint violation.

Here, $\eta_\phi < \eta_\theta < \eta_\lambda$ are learning rates, ensuring the updates occur at different timescales.

In power system applications, Lagrangian-based methods have been effectively implemented across various control challenges. Wang *et al.* (2019) apply PPO-Lagrangian to Volt-VAR control, distinctively incorporating voltage constraints directly into the Lagrangian formulation for distribution systems. Zhang *et al.* (2020b) use explicit information from the power flow equations and operational constraints to obtain gradients (2.2) and develop a distributed consensus-based training algorithm where agents coordinate through Lagrangian multipliers, rather than using centralized training as in standard CPO/TRPO. Hu *et al.* (2024) develop the PDTD3 algorithm through RCPO, achieving near-optimal performance with millisecond-level computation times while managing both single-step and time-coupling constraints. The success of Lagrangian methods in these applications stems from their ability to handle multiple constraints simultaneously while maintaining computational efficiency—particularly valuable for real-time power system operations.

### 2.1.3 Algorithm-Specific Extensions: Soft Actor-Critic (SAC)

#### Overview of Standard SAC

The standard SAC algorithm incorporates three key ingredients: an actor-critic architecture with separate policy and value function networks, an off-policy formulation that enables reuse of previously collected data for efficiency, and entropy maximization to enable stability and exploration. The objective incorporates both the expected return and the entropy of the policy (Haarnoja *et al.*, 2018):

$$J(\pi) = \sum_t \mathbb{E}_{(s_t,a_t)\sim d^\pi}[r(s_t,a_t) + \tau H(\pi(\cdot|s_t))] \tag{2.8}$$

where $d^\pi(s_t,a_t)$ is the state-action marginal of the trajectory distribution, $H(\pi(\cdot|s_t))$ is the entropy of the policy at state $s_t$, and $\tau$ is a temperature parameter balancing exploitation and exploration.

SAC employs a parametric state value function $V_{\phi_1}(s_t)$ and soft Q-function $Q_{\phi_2}(s_t,a_t)$, and a policy network $\pi_\theta(a_t|s_t)$. While the state value function can be estimated from a single action sample from the Q-function and policy, introducing separate function approximator for the soft value can stabilize training and is convenient to train simultaneously with the other networks. The updates for these networks are:

1. Critic update: The state value function is trained to minimize the squared residual error

$$J_V(\phi_1) = \mathbb{E}_{s_t\sim\mathcal{D}}\left[\frac{1}{2}\left(V_{\phi_1}(s_t) - \mathbb{E}_{a_t\sim\pi_\theta}(Q_{\phi_2}(s_t,a_t) - \log\pi_\theta(a_t|s_t))\right)^2\right]$$

where $\mathcal{D}$ is the distribution of previously sampled states and actions, or a replay buffer. The soft Q-function is updated to minimize the temporal difference error, a.k.a., Bellman residual:

$$J_Q(\phi_2) = \mathbb{E}_{(s_t,a_t)\sim\mathcal{D}}\left[\frac{1}{2}(Q_{\phi_2}(s_t,a_t) - (r(s_t,a_t) + \gamma\mathbb{E}_{s_{t+1}}V_{\bar{\phi}_1}(s_{t+1})))^2\right]. \tag{2.9}$$

Here, $V_{\bar{\phi}_1}$ is a target value network, where $\bar{\phi}_1$ can be an exponentially moving average of the value network weights to stabilize training.

2. Policy update: The policy is updated to minimize the expected KL-divergence with the Q-value network:

$$J_\pi(\theta) = \mathbb{E}_{s_t\sim\mathcal{D}}\left[\mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t)}\left[\log\pi_\theta(a_t|s_t) - \log\frac{\exp(Q_{\phi_2}(s_t,a_t))}{Z_{\phi_2}(s_t)}\right]\right]$$

where $Z_{\phi_2}(s_t)$ is the normalization factor. We can apply the reparametrization trick[1] by expressing the action as a deterministic function of a

[1] The reparametrization trick enables computing gradients of expectations by transforming a sampling operation into a deterministic function of a fixed noise

spherical Gaussian: $a_t = a_\theta(\varepsilon_t; s_t)$ with $\varepsilon_t \sim \mathcal{N}$ denotes sampling from a standard normal distribution. This leads to the equivalent objective (after dropping the participation term $Z_{\phi_2}(s_t)$ that does not depend on $\theta$:

$$J_\pi(\theta) = \mathbb{E}_{s_t \sim \mathcal{D}, \varepsilon_t \sim \mathcal{N}}[\log \pi_\theta(a_\theta(\varepsilon_t; s_t)|s_t) - Q_{\phi_2}(s_t, a_\theta(\varepsilon_t; s_t))],$$

where $\pi_\theta$ is defined implicitly in terms of $a_\theta$.

This transformed objective can now be optimized using standard backpropagation through both the policy network ($f_\theta$) and the Q-function network ($Q_{\phi_2}$), providing lower-variance gradient estimates compared to the REINFORCE estimator (Williams, 1992).

## Constrained Soft Actor-Critic (CSAC)

CSAC, or primal-dual SAC, combines SAC with primal-dual methods to handle constraints (see, e.g., (Zhang *et al.*, 2023d)):

$$\begin{aligned} \max_\pi J(\pi) &= \sum_t \mathbb{E}_{(s_t, a_t) \sim d^\pi}[r(s_t, a_t) + \tau H(\pi(\cdot|s_t))] \\ \text{subject to } J_c(\pi) &= \sum_t \mathbb{E}_{(s_t, a_t) \sim d^\pi}[c(s_t, a_t)] \le \xi, \end{aligned}$$

where $c(s_t, a_t)$ is the constraint cost for a state-action pair and $\xi$ is the constraint threshold. CSAC introduces a cost critic $Q^c_\psi(s, a)$ alongside two Q-functions $Q_{\phi_1}(s, a)$ and $Q_{\phi_2}(s, a)$ to reduce overestimation bias. The policy update becomes:

$$J_\pi(\theta) = \mathbb{E}_{s \sim \mathcal{D}, \varepsilon \sim \mathcal{N}}[\tau H(\pi_\theta(\cdot|s)) - \min_{i=1,2} Q_{\phi_i}(s, a_\theta(\varepsilon; s)) + \lambda Q^c_\psi(s, a_\theta(\varepsilon; s))]$$

distribution. Consider computing: $\nabla_\theta \mathbb{E}_{q_\theta(z)}[f(z)]$, where $f(z)$ is some function and $z$ is sampled from a distribution $q_\theta(z)$ with parameters $\theta$. The key insight is expressing $z$ as a deterministic transformation of a parameter-free random variable $\varepsilon$ (e.g., $p$ can be the normal distribution):

$$z = g_\theta(\varepsilon), \quad \varepsilon \sim p(\varepsilon)$$

This allows rewriting the gradient as:

$$\nabla_\theta \mathbb{E}_{p(\varepsilon)}[f(g_\theta(\varepsilon))] = \mathbb{E}_{p(\varepsilon)}[\nabla_\theta f(g_\theta(\varepsilon))]$$

This formulation contrasts with the REINFORCE estimator, which uses $\mathbb{E}_{q_\theta(z)}[f(z)\nabla_\theta \log q_\theta(z)]$. While REINFORCE works for both discrete and continuous variables, the reparametrization trick typically provides lower variance gradient estimates for continuous variables, making it the preferred choice when applicable.

where $a_\theta(\varepsilon; s)$ is the reparameterized action and $\lambda$ is a Lagrange multiplier. This equation updates the policy to maximize reward and entropy while minimizing constraint violations. The Lagrange multiplier is updated as:

$$\lambda_{i+1} = [\lambda_i + \eta_\lambda(J_c(\pi_{\theta_i}) - \xi)]_+ \tag{2.10}$$

where $\eta_\lambda$ is the learning rate for the Lagrange multiplier.

The primal-dual formulation enables explicit handling of operational constraints through cost critics, while the entropy regularization promotes exploration of safe operating regions. For example, Zhang *et al.* (2023d) applied CSAC to EV charging by designing the cost function $c(s_t, a_t)$ that combines both BES operational limits and EV charging requirements. The action space design reduces dimensionality by grouping EVs into sets based on charging states, transforming an $N$-dimensional control problem into a two-dimensional one: BES operation and aggregate charging power. A safety filter validates control actions before execution, ensuring constraint satisfaction during both training and deployment. This implementation shows how CSAC can be adapted to handle specific microgrid operation challenges while maintaining computational tractability.

### Risk-aware Soft Actor-Critic (RSAC)

RSAC incorporates risk-awareness using Conditional Value-at-Risk (CVaR):

$$\Gamma_\pi(s, a, \beta) \leq \xi, \tag{2.11}$$

were $\Gamma_\pi(s, a, \beta)$ is the CVaR of the cumulative cost distribution, $\beta$ is the risk level that can be tuned based on system requirements, and $\xi$ is the CVaR threshold that enforces constraint satisfaction. For a Gaussian distribution, the CVaR is calculated as:

$$\Gamma_\pi(s, a, \beta) = Q_c^\pi(s, a) + \beta^{-1}\zeta(\mathcal{Z}^{-1}(\beta))\sqrt{Var_c^\pi(s, a)} \tag{2.12}$$

where $Q_c^\pi(s, a)$ represents the expected cumulative cost, $Var_c^\pi(s, a)$ captures its variance, $\zeta(\cdot)$ and $\mathcal{Z}^{-1}(\cdot)$ are the probability density function and the inverse cumulative distribution function of the standard normal distribution.

The policy update is governed by:

$$J_\pi(\theta) = -\mathbb{E}_{s_t \sim \mathcal{D}, a_t \sim \pi}[Q_r^\pi(s_t, a_t) - \tau \log \pi(a_t|s_t) - \lambda \Gamma_\pi(s_t, a_t, \beta)] \quad (2.13)$$

where $Q_r^\pi$ represents the reward-critic, $\tau$ controls exploration through entropy regularization, and $\lambda$ weights the importance of constraint satisfaction through the CVaR term. This formulation balances the competing objectives of maximizing performance while maintaining safety constraints.

Through this formulation, RSAC effectively handles the stochastic nature of renewable generation while maintaining strict operational bounds on critical system parameters. For example, in (Yu *et al.*, 2024), RSAC is deployed to control a district cooling system's mass flow rate based on tie-line power and temperature measurements, optimizing power smoothing while maintaining building temperatures with probabilistic guarantees. This essentially creates grid-scale energy storage from building thermal mass while ensuring occupant comfort under uncertainty.

### Human-Guided Safe SAC

The Human-Guided Safe SAC aims to incorporate human expertise into the SAC algorithm, particularly for power systems control (Sun *et al.*, 2024). At each timestep, the framework dynamically chooses between the learned policy and human guidance:

$$a_t = I(s_t) \cdot a_t^{HM} + [1 - I(s_t)] \cdot a_t^{DRL},$$

where $I(s_t)$ is a binary indicator function that evaluates the current state $s_t$ and returns 1 if human intervention is needed (e.g., when voltage violations are detected) and 0 otherwise, $a_t^{HM}$ represents the human-guided action, while $a_t^{DRL}$ is the action proposed by the learned policy. A hybrid experience replay buffer that stores both regular transitions and transitions where human intervention was needed. When voltage violations occur, the buffer stores both the original unsafe transition with a modified reward that includes a penalty, and the safe transition after human intervention with the original reward.

The policy is updated by minimizing a loss function that encourages the policy to learn from human-guided actions while still maximizing expected returns and maintaining high entropy:

$$J_\pi(\theta) = \mathbb{E}_{\mathcal{D}}[\tau \log \pi_\theta(a_{t+1}|s_{t+1}) - Q_\phi(s_t, a_t) + \omega_I \|\pi_\theta(a_{t+1}|s_{t+1}) - a_t\|_2^2], \tag{2.14}$$

where $\tau \log \pi_\theta(a_{t+1}|s_{t+1})$ encourages exploration through entropy maximization and $\omega_I \|\pi_\theta(a_{t+1}|s_{t+1}) - a_t\|_2^2$ encourages the policy to learn from both DRL and human-guided actions, with $\omega_I > 0$ denoting the human intervention factor.

The critic networks are trained to minimize the soft Bellman residual (2.9). Following SAC, two critic networks are employed to prevent value overestimation, with the minimum Q-value used for policy updates.

Sun *et al.* (2024) studied volt/var control of photovoltaic inverters, where human expertise guides the sequential correction of voltage violations and phase unbalances. Voltage sensitivity-based guidance rules are employed to encode domain knowledge about how reactive power adjustments affect voltage profiles in distribution networks. This domain-specific guidance helps maintain stability during both training and deployment while allowing the SAC framework to discover optimal policies that minimize power losses.

### 2.1.4 Strengths and Limitations of Different Approaches

Trust region methods (e.g., CPO and PCPO) provide robust theoretical guarantees and stable learning for power system control through carefully bounded policy updates, though their computational demands can limit real-time applications in large networks. Dynamic switching approaches (CRPO, PCRPO, ESPO) offer more straightforward implementation with better sample efficiency, but may struggle with multiple constraints.

Primal-dual methods, particularly PPO-Lagrangian and TRPO-Lagrangian, have shown promise in handling multiple constraints simultaneously, such as economic efficiency (e.g., minimizing generation costs) and system security (e.g., maintaining adequate reserves, voltage/frequency regulation). Compared to primal-based methods, the constraints appear additively in the Lagrangian function, essentially reducing to a single objective. However, these methods often require careful tuning of learning parameters and potentially suffer from temporary constraint violations during training. RCPO introduces additional stability through multiple timescales, valuable for power systems with varying operational dynamics, though at the cost of increased complexity and slower convergence.

Whitin the SAC-based family of methods, CSAC enables efficient off-policy learning for grid control but may struggle with hard operational constraints, while RSAC explicitly handles the uncertainty inherent in renewable generation through CVaR, though it may be overly conservative in maintaining grid stability. Human-guided Safe SAC leverages valuable power system operator expertise but depends heavily on the availability of expert knowledge and may not generalize well to unprecedented grid conditions. The choice of method for power system applications ultimately depends on specific requirements: trust region methods for strict operational constraints, primal-dual methods for managing multiple competing grid objectives, risk-aware methods for handling renewable uncertainty, and human-guided approaches for scenarios where operator expertise can be effectively codified.

## 2.2 Design and System Architecture Elements

This section bridges theoretical methods with practical power system requirements, examining implementation approaches that have shown promise in maintaining system safety.

### Constraint Formulation and Handling

Effective constraint formulation and handling are crucial for safe RL in power systems. Domain-specific constraints can be incorporated directly into the RL formulation. For example:, voltage limits: $v_{\min} \leq v_j \leq v_{\max}, \forall j \in \mathcal{N}$, and line thermal constraints: $|p_{jj'}| \leq p_{jj'}^{\max}, \forall (j, j') \in \mathcal{E}$, where $\mathcal{N}$ is the set of buses and $\mathcal{E}$ is the set of lines. These constraints can be incorporated into the reward function or handled explicitly in constrained RL formulations. CVaR-based approaches, such as RSAC, incorporate risk awareness into the constraint formulation, which is particularly useful for handling uncertainties in renewable energy integration and demand forecasting.

Safety state augmentation, as exemplified by Sauté MDP (Sootla *et al.*, 2022), transforms the constrained MDP into an unconstrained MDP with an augmented state space, allowing standard RL algorithms to implicitly handle safety constraints. In this method, the state space is augmented with a safety budget:

$$s'_t = [s_t, z_t],$$

where $z_t$ is the safety budget, initialized as the initial safety threshold $z_0 = \xi$ and updated as:

$$z_{t+1} = (z_t - c(s_t, a_t))/\gamma.$$

One advantage of this method is the plug-and-play nature, i.e., any RL algorithm can be "Sautéed." Also, state augmentation allows for policy generalization across safety constraints, since the threshold is now part of the state rather than the CMDP formulation.

### Action Space Design

Proper action space design is essential for effective RL in power systems. The action space typically contains both discrete and continuous control variables. For instance, in the problem of optimal operation of distribution networks (Li and He, 2022), switchable capacitor banks (SCBs), the tap position of the on-load tap-changers (OLTCs) and voltage regulators (VRs) operate in discrete steps whereas dispatchable generators and battery storage systems operate with continuous outputs. To deal with the mixed discrete and continuous action space, a typical approach is to approximate the policy by using a joint distribution:

$$\pi_\theta(a_t|s_t) = \pi^c_\theta(a^c_t|s_t) \cdot \pi^d_\theta(a^d_t|s_t), \tag{2.15}$$

with $\pi^c_\theta(a^c_t|s_t)$ and $\pi^d_\theta(a^d_t|s_t)$ capturing the continuous and discrete parts of the policy, respectively.

For discrete actions, we can use the softmax function as the probability distribution:

$$\pi^d_\theta(a^d = a|s) = \frac{\exp(f^d_\theta(s)_a)}{\sum_{a'} \exp(f^d_\theta(s)_{a'})} \tag{2.16}$$

where $f_\theta^d(s)_a$ is the $a$-th output of the discrete action network.

For continuous actions, we typically use Gaussian distribution with parametrized mean function and some specification of the variance. Also, discretization can sometimes simplify the action space:

$$a_t^d = \text{round}(a_t^c/\delta_{\text{disc}}) \cdot \delta_{\text{disc}} \tag{2.17}$$

where $\delta_{\text{disc}}$ is the discretization step. This approach can be particularly useful for actions like transformer tap changes or capacitor bank switching.

### Exploration Strategies

Safe exploration is critical in power system applications to prevent dangerous or costly actions during learning.Entropy-regularized exploration can be modified to account for safety, e.g., CSAC, where safety cost is combined with the reward. Another approach to enforcing hard constraints is to use a safety layer that projects unsafe actions onto the safe action space:

$$a_t^{\text{safe}} = \mathcal{P}_{\mathcal{A}_{\text{safe}}}(a_t)$$

where $\mathcal{A}_{\text{safe}}$ is the set of safe actions. This can be implemented using techniques like constrained optimization or barrier functions. Chapter 5 provides a deep dive into this topic.

## 2.3 System-Level Considerations and Future Directions

Power system applications of SRL face several interconnected challenges spanning implementation, scalability, and adaptation. The integration of safety constraints with performance optimization remains a central concern, where different methods offer distinct trade-offs.

**Scalability and Computational Efficiency** Scalability presents a critical challenge as power systems grow in complexity. Distributed learning approaches, particularly those using parallel actors and shared policy networks, have shown significant promise. Distributed Actor-Critic methods scale up learning through parallel experience collection, with gradient updates following:

$$\nabla_\theta J(\theta) \approx \frac{1}{K} \sum_{k=1}^{K} \nabla_\theta J_k(\theta),$$

where $K$ parallel actors share policy parameters $\theta$, and $J_k$ is the policy objective for actor $k$. Zhang *et al.* (2023a) demonstrated this approach's effectiveness through Distributed PPO for EV charging coordination, achieving higher sample efficiency compared to a single-actor PPO and faster convergence in high-dimensional EV allocation task.

Multi-Agent Reinforcement Learning (MARL) offers another scaling approach by decomposing systems into interacting agents, each controlling a distinct subsystem with local policy $\pi_k(a_k|s_k)$ (Chen *et al.*, 2021). The key challenge becomes balancing local optimization with global system stability, often addressed through techniques like networked value functions that incorporate neighboring agents' states (Chen *et al.*, 2021), or consensus algorithms for policy coordination (Fan *et al.*, 2023b). This paradigm handles heterogeneous agents (e.g., different types of power demands/supplies) and varying coordination levels, particularly suitable for modern power systems with diverse distributed resources. See Chapter 4 for a comprehensive discussion.

**Adaptation to Changing Environments** Adaptation to changing system conditions is particularly relevant in modern power grids, demanding safe RL algorithms that can quickly adapt to evolving grid topologies, generation mixes, and load patterns. Meta-Safe RL (Meta-SRL) addresses this challenge by framing power system control as a sequence of CMDPs, where each task $m$ aims to maximize expected reward $J_{m,r}(\pi)$ subject to constraints $J_{m,c}(\pi) \leq \xi_m$ (see Figure 2.2 for the framework overview). At the base level, an SRL algorithm, such as CRPO, learns control policies for specific grid conditions by alternating between reward maximization and constraint satisfaction, following the gradient updates we saw earlier: $\pi_{m,i+1} = \pi_{m,i} + \eta_m g_m^r$ or $\pi_{m,i} + \eta_m g_m^c$, where $g_m^r$ and $g_m^c$ are the reward and cost gradients as defined in (2.2). The key innovation comes at the meta level, where online learning algorithms adapt initialization parameters $(\pi_{m,0}, \eta_m)$ across tasks to minimize both the task-averaged regret $\bar{R}_r = \frac{1}{M}\sum_{m=1}^{M}(J_{m,r}(\pi_m^*) - \mathbb{E}[J_{m,r}(\hat{\pi}_m)])$ and

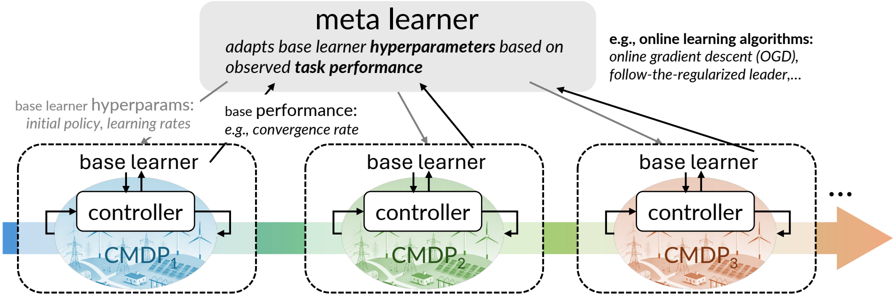

meta learner
adapts base learner hyperparameters based on observed task performance
e.g., online learning algorithms:
online gradient descent (OGD), follow-the-regularized leader,...
base learner hyperparams:
initial policy, learning rates
base performance:
e.g., convergence rate
base learner
controller
$CMDP_1$
base learner
controller
$CMDP_2$
base learner
controller
$CMDP_3$
...

**Figure 2.2: Meta-SRL framework.** A meta-learner employs online learning algorithms (e.g., OGD) to adapt base learner hyperparameters (initial policy $\pi_{m,0}$ and learning rates $\eta_m$) across a sequence of CMDPs. For each CMDP task, the base learner (implemented as CRPO) learns a control policy while balancing reward maximization and constraint satisfaction. The meta-learner optimizes these hyperparameters based on the base learner's performance metrics, including convergence rates and constraint violations, enabling efficient transfer across related constrained control tasks (Khattar *et al.*, 2023).

constraint violations $\bar{R}_c = \frac{1}{M}\sum_{m=1}^{M}(\xi_m - J_{m,c}(\hat{\pi}_m))$, where $\pi_m^*$ is the optimal policy and $\hat{\pi}_m$ is the learned policy. This adaptation leverages task similarity measured via KL divergence between optimal policies for different grid states. The framework shows theoretical convergence rates that is proportional to the similarity between grid operating conditions.

Beyond current safe RL methods which typically treat power system control as a single CMDP, Meta-SRL offers a systematic way to handle temporal evolution of grid conditions while maintaining safety guarantees. This opens new possibilities for developing adaptive control strategies that can efficiently transfer knowledge between different operating scenarios while ensuring reliable grid operation. Further research is needed to validate these theoretical guarantees in realistic power system environments and develop computationally efficient implementations suitable for real-time grid control.

# 3

# Safe Model-Based RL

Models, much like maps, are inevitably partial representations that must still guide effective decision-making. In safe model-based reinforcement learning (MBRL), these representations must serve multiple purposes while acknowledging their inherent limitations.

Safe model-based RL leverages a learned dynamics model, represented as $\dot{s} = f(s, a, w)$ (where $w$ captures noise or unmodeled information), to make informed decisions about safety and optimization. This approach offers several unique advantages. First, it enhances sample efficiency by allowing the agent to reason about consequences without direct experience. Second, it provides interpretability through the learned model, offering insights into system behavior. Third, it potentially allows for transferability, as a learned model can often generalize to new tasks.

The framework of safe MBRL revolves around four key components:

- Model Learning: This component focuses on learning the system dynamics $f$ from safely collected data. Common approaches include Gaussian Processes, Neural Networks, and Ensemble methods. Advanced learning approaches enhance this component through joint Learning of models and certificates for end-to-end

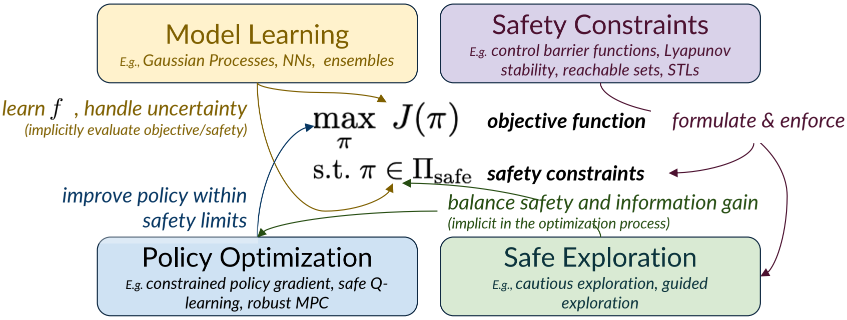

Model Learning
E.g., Gaussian Processes, NNs, ensembles
Safety Constraints
E.g. control barrier functions, Lyapunov stability, reachable sets, STLs
learn $f$, handle uncertainty
(implicitly evaluate objective/safety)
$\max_{\pi} J(\pi)$
objective function
formulate & enforce
s.t. $\pi \in \Pi_{\text{safe}}$
safety constraints
improve policy within safety limits
balance safety and information gain
(implicit in the optimization process)
Policy Optimization
E.g. constrained policy gradient, safe Q-learning, robust MPC
Safe Exploration
E.g., cautious exploration, guided exploration

**Figure 3.1: Overview of safe MBRL.** The core components (model learning, safety constraints, safe exploration, and policy optimization) form a cyclic learning process centered around a constrained optimization problem $\max_{\pi} J(\pi)$ subject to $\pi \in \Pi_{\text{safe}}$. The framework is enhanced by learning approaches (joint learning, certificate learning, and uncertainty quantification) that enable data-efficient safety verification, and practical considerations (structure-constrained learning, computational efficiency, and generalization) that facilitate real-world implementation. Arrows indicate information flow and functional relationships between components, showing how safety constraints and performance objectives are balanced throughout the learning process.

training.

- Safety Constraints: Here, the goal is to formulate and enforce safety constraints (Chapter 1.4). Approaches include control-theoretic tools such as Control Barrier Functions, Lyapunov Stability analysis, and Reachable Set computation.

- Safe Exploration: This involves collecting data while ensuring the policy remains within a safe set. Strategies include Optimistic, Cautious, and Guided Exploration.

- Policy Optimization: The objective is to improve the policy $\pi$ while maintaining safety. Techniques include Constrained Policy Gradient methods, Safe Q-learning, and Robust Model Predictive Control.

These four components are deeply interconnected and enhanced by practical considerations for real-world implementation (Figure 3.1 provides an overview). For instance, a more accurate model enables more precise safety constraints and safer exploration, while structure-constrained learning incorporates domain knowledge, computational efficiency enables scalability, and generalization approaches ensure robust performance across different operating conditions.

The field of safe model-based RL faces several key challenges: **1)** Handling model uncertainty: As the learned model is an approximation, managing and accounting for model errors is crucial for safety guarantees. **2)** Balancing exploration and safety: There's a fundamental tension between the need to explore for better learning and the requirement to maintain safety at all times. **3)** Ensuring constraint satisfaction under uncertainty: Safety constraints must be satisfied not just for the nominal model, but for all plausible models within the uncertainty set. **4)** Scalability to complex systems: As system complexity increases, computational demands for model learning, constraint evaluation, and policy optimization can become prohibitive. **5)** Providing formal safety guarantees: Establishing rigorous safety proofs in the presence of learned components and uncertainties remains a significant challenge.

**Notion Primer** This chapter uses both continuous-time dynamics $\dot{s} = f(s, a)$ with trajectories $s(t)$, and discrete-time dynamics $s_{t+1} = f(s_t, a_t)$ with sequences $\{s_t\}_{t=0}^{T}$, where $s \in \mathcal{S}$ represents system states and $a \in \mathcal{A}$ denotes control actions. For systems with uncertainty, we write $f(s, a, w)$ where $w \in \mathcal{W}$. Safety certificates use Lie derivatives $L_f h(s) = \nabla h(s)^\top f(s)$ in continuous time and forward differences $\Delta h(s) = h(f(s)) - h(s)$ in discrete time. Safety constraints are defined through sets $\mathcal{S}_{\text{safe}} = \{s \mid h(s) \geq 0\}$, while stability is analyzed using Lyapunov functions $\Phi(s)$ and barrier functions $B(s)$. Power system variables include voltage $v_j$ at bus $j$ with bounds $[\underline{v}_j, \bar{v}_j]$, frequency deviation $\Delta f_k$, and Area Control Error $ACE_k$ in area $k$. Learning components include parameterized policies $\pi_\theta(s)$, neural certificates with parameters $\psi$, value functions $V^\pi(s)$, and respective loss functions $\mathcal{L}_{\text{cert}}$ and $\mathcal{L}_{\text{policy}}$ for optimization.

## 3.1 Control-Theoretic Approaches

Control-theoretic approaches aim to combine the data-efficiency of model-based methods with rigorous safety guarantees, enabling the application of RL in safety-critical domains.

Lyapunov functions are fundamental tools for proving stability of nonlinear systems.

**Definition 3.1** (Lyapunov Function). For a system $\dot{s} = f(s)$, a continuously differentiable function $\Phi : \mathbb{R}^n \to \mathbb{R}$ is a Lyapunov function if: (i) $\Phi(s) > 0$ for all $s \neq 0$, $\Phi(0) = 0$ (ii) $\dot{\Phi}(s) = L_f\Phi(s) = \nabla\Phi(s)^\top f(s) < 0$ for all $s \neq 0$.

They provide a way to show that the system's "energy", in some generalized sense, is always decreasing. This stability definition is typically seen in frequency control, where convergence to a synchronized state where all frequencies match and phase angles settle to values satisfying the power flow equations is desirable (Cui *et al.*, 2023; Liu *et al.*, 2024b).

For voltage control problems, we often need to prove convergence to a safe operating band $S_v = \{v \in \mathbb{R}^{|\mathcal{N}|} : \underline{v}_j \leq v_j \leq \bar{v}_j, j \in \mathcal{N}\}$ rather than to a specific equilibrium point. LaSalle's invariance theorem extends Lyapunov theory to handle such cases by showing convergence to the largest invariant set where the Lyapunov function derivative is zero. This provides a natural framework for analyzing set-based stability without requiring the Lyapunov function to be zero only at a single point. Feng *et al.* (2023) leverage this principle for discrete-time voltage control dynamics. By designing the controller to be zero inside $S_v$, they ensure this set is invariant, thus guaranteeing voltage stability through LaSalle's theorem while accommodating the practical requirement of maintaining voltages within an acceptable range rather than at a fixed point.

Control Lyapunov Functions (CLF extend the concept of Lyapunov functions to controlled systems, providing a constructive way to design stabilizing controllers.

**Theorem 3.1** (Control Lyapunov Function). For a system $\dot{s} = f(s, a)$, if there exists a continuously differentiable function $\Phi : \mathbb{R}^n \to \mathbb{R}$ such that: (i) $\Phi(s) > 0$ for all $s \neq 0$, $\Phi(0) = 0$ (ii) $\inf_a\{\nabla\Phi(s)^\top f(s, a)\} < 0$ for all $s \neq 0$ Then there exists a stabilizing control law.

Barrier functions provide a way to verify set invariance, another type of safety constraint.

**Definition 3.2** (Barrier Function). For a safe set $\mathcal{S}_{\text{safe}} = \{s \in \mathbb{R}^n : h(s) \geq 0\}$, a continuously differentiable function $B : \mathbb{R}^n \to \mathbb{R}$ is a barrier function if: $\dot{B}(s) \geq -\alpha(B(s))$ for all $s \in \mathcal{S}_{\text{safe}}$ where $\alpha$ is a class $\mathcal{K}$ function.[1]

Control Barrier Functions (CBFs) extend barrier functions to controlled systems, providing a way to design controllers that ensure safety. Some specific types of CBFs can be obtained by choosing $B(\cdot)$ as the safety constraint $h(\cdot)$ (zeroing CBF) or its reciprocal (reciprocal CBF) to provide additional flexibility in shaping the barrier function behavior near the boundary of the safe set. For instance, AdaSafe (Wan *et al.*, 2023) use both for the frequency control problem.

**Theorem 3.2** (Control Barrier Function). For a system $\dot{s} = f(s, a)$ and safe set $\mathcal{S}_{\text{safe}} = \{s \in \mathbb{R}^n : h(s) \geq 0\}$, if there exists a continuously differentiable function $B : \mathbb{R}^n \to \mathbb{R}$ such that: $\sup_a\{\nabla B(s)^\top f(s, a) + \alpha(B(s))\} \geq 0$ for all $s \in \mathcal{S}_{\text{safe}}$ where $\alpha$ is a class $\mathcal{K}$ function, then there exists a control policy that renders $\mathcal{S}_{\text{safe}}$ forward invariant.

Zhao *et al.* (2023) implement CBFs to ensure safe operation of synchronous generators and inverter-based distributed generators. Specifically, they modify unsafe RL control actions through an optimization problem $\phi(u_r) = \max(0, \nabla B(s)^\top f(s, a^{\text{rl}} + u_r) + \alpha(B(s))) + \mu\|u_r\|_2^2$. Here, $a^{\text{rl}}$ is the RL action and $\mu$ penalizes large modifications. The safety filter updates the refined control $u_r' = u_r - \nabla_{u_r}\phi(u_r)$ until CBF conditions are met, ensuring operational constraints during transients.

Robust Control Lyapunov Barrier Functions (rCLBFs) integrate stability requirements, as expressed by CLFs, with safety constraints, as captured by CBFs, while accounting for model uncertainties. Wang *et al.* (2023a) use the rCLBF approach to encode both the requirement that the system states converge to a stable equilibrium and the condition that they remain within safe voltage and frequency bounds, even under uncertain and time-varying renewable generation. The practical benefit

[1] A class $\mathcal{K}$ function is a continuous, strictly increasing function $\alpha : [0, \infty) \to [0, \infty)$ with $\alpha(0) = 0$. These functions are commonly used in control theory and stability analysis to define comparison functions and characterize rates of convergence. A simple example would be $\alpha(x) = x$ and $\alpha(x) = x^2$ for $x \geq 0$.

is that the resulting controller can handle a wide range of disturbances while guaranteeing both safety and stability, thereby increasing the reliability and flexibility of power system operations without sacrificing performance.

Robust control methods provide systematic frameworks for handling model uncertainties and disturbances in control system design (Zhou and Doyle, 1998). These approaches explicitly account for the gap between mathematical models and physical systems, ensuring stability and performance despite uncertainties.

One powerful framework within robust control is Integral Quadratic Constraints (IQCs), which provide a general approach to characterize the input-output behavior of uncertain or nonlinear operators in closed-loop systems (Megretski and Rantzer, 1997). IQCs unify various notions of stability and performance by expressing conditions as quadratic forms integrated over time. Before defining IQCs, we establish a few notations. Let $L_2^n[0,\infty)$ denote the space of square-integrable signals $y : [0,\infty) \to \mathbb{R}^n$, with the $L_2$ norm given by $\|y\|_{L_2} = \left(\int_0^\infty \|y(t)\|^2 dt\right)^{1/2}$. A causal operator $\Delta : L_2^n[0,\infty) \to L_2^m[0,\infty)$ maps an input signal $y(t)$ to an output $w(t) = \Delta(y)(t)$. Below, we introduce the time-domain definition of IQC (Seiler, 2014).

**Definition 3.3** (Time-domain IQC)**.** Let $\Delta : L_2^n[0,\infty) \to L_2^m[0,\infty)$ be a bounded, causal operator. Consider a stable, linear operator $\Psi : L_2^{n+m}[0,\infty) \to L_2^q[0,\infty)$ and a constant, symmetric matrix $M \in \mathbb{R}^{q\times q}$. Define the filtered signals $z(t) = \Psi \begin{bmatrix} y(t) \\ w(t) \end{bmatrix}$, where $w(t) = \Delta(y)(t)$. We say that $\Delta$ satisfies the IQC defined by $(\Psi, M)$ if for all $y \in L_2^n[0,\infty)$,

$$\int_0^\infty z(t)^\top M z(t)\, dt \geq 0.$$

Intuitively, this inequality restricts the way $w = \Delta(y)$ can deviate from $y$ by enforcing a quadratic energy-like bound. Various known uncertainty classes (e.g., sector-bounded nonlinearities, slope-restricted functions) can be encoded as IQCs with suitable choices of $\Psi$ and $M$. IQCs have been applied to analyze the stability of power system controllers under large uncertainties. For example, Jin and Lavaei (2020) use IQCs, combined with Lyapunov and dissipativity-based methods, to certify stability of RL policies applied to nonlinear and uncertain power network models. By treating the RL-based controller and nonlinear power flow dynamics as interconnected blocks, IQC analysis provides a systematic way to ensure that, despite model inaccuracies and changing operating conditions, the closed-loop system will remain stable. This complements the earlier introduced Lyapunov and barrier function techniques by offering a scalable and flexible framework to handle complex, data-driven policies in real-world power systems.

Several other mathematical frameworks complement these approaches for analyzing stability, safety, and robustness. Contraction metrics (Lohmiller and Slotine, 1998) provide a differential framework for analyzing trajectory convergence by studying how infinitesimal distances between solutions evolve over time, offering tools for studying nonlinear system stability and robustness. Reachability analysis provides formal methods to compute sets of states that can be reached under bounded disturbances, with Hamilton-Jacobi methods being particularly useful for safety verification (c.f., (Bansal *et al.*, 2017) for a recent overview). Dissipativity theory (Willems, 1972) generalizes energy-based notions such as passivity to characterize input-output properties of dynamical systems, with recent applications to networked systems (Arcak *et al.*, 2018). Together, these tools offer different perspectives and techniques for ensuring safe control of power systems (See Table 3.1 for an overview).

## 3.2 Learning-based Certificates and Joint Policy Learning

To build upon the previously introduced concepts, we now present a generic recipe for integrating certificates (e.g., Lyapunov functions, CBFs) with the policy optimization process in safe MBRL (Fig. 3.2). This recipe provides a systematic approach to ensure that safety constraints and performance objectives are jointly addressed, enabling a principled pathway toward end-to-end learning of safe and effective control policies.

The process begins with access to a dataset of system trajectories, which may be obtained from historical records or simulation. A suitable model of the system dynamics, assumed to be given or known, provides

**Table 3.1:** Comparison of three RL approaches for power system control, analyzing their problem formulation, theoretical stability guarantees, learning methodology, and implementation considerations.

| **Problem Setting** | **Safety Guarantee** | **Learning Approach** | **Practical Aspects** |
|---|---|---|---|
| (Jin and Lavaei, 2020) Frequency regulation in power system; Mixed linear-nonlinear dynamics; Uncertainties from renewables, loads, parameter variations; Unknown nonlinear components | IQC-based certificates; Robust invariance sets under bounded uncertainties; Requires Hurwitz nominal system | TRPO with stability regularization; Neural nets with bounded gradients; Safe exploration guided by IQC certificates | Requires solving SDPs offline; Safe exploration and training; Local measurements; Handles nonlinear uncertainties |
| (Gu *et al.*, 2022) Frequency regulation in multi-area system; Partial observability; Both linear and nonlinear dynamics; Parameter uncertainties considered | Exponential stability via IQCs; Lyapunov theory with S-lemma; Sequential convexification | Policy gradient with stability projection; RNN architecture with tanh activation; Convex stability conditions | Requires solving SDPs; Safety during exploration and training; Handles decentralized control |
| (Cui *et al.*, 2023) Primary frequency control; Renewable integration; Nonlinear swing equations; System parameter/initial condition variations | Local exponential stability; Assumes phase angles in $[0, \pi/2)$; Lyapunov construction | ReLU-based RNN with saturation and monotonicity constraints; Stability by design | Local measurements; No runtime optimization; Compatible with standard inverter interfaces |
| (Feng *et al.*, 2023) Voltage control in distribution systems; Local measurements only; Discrete-time dynamics; Exact network parameters unknown | LaSalle's invariance; Assumes network reactance matrix is positive definite | Modified DDPG; Monotone neural nets; ReLU architecture; Zero control in safe voltage band | <1ms computation; Scales to 123-bus; Local measurements; Real-time capable |
| (Liu *et al.*, 2024b) Load frequency control in multi-area system; Maintains stability under adversarial attacks on communication network | Lyapunov stability; Empirical robustness against FGSM attacks; Monotonic policy | DDPG + adversarial training; Split positive/negative ReLU nets; Monotonicity constraints | Faster convergence (than standard DDPG); Requires communication; Handles attack scenarios |
| (Zhao *et al.*, 2023) Transient stability and voltage control; Nonlinear networked power systems; Both synchronous and inverter-based generators; No explicit uncertainty modeling | Forward invariance via control barrier functions; Requires accurate state measurements; Deterministic guarantees | DDPG with neural barrier certification; Two-layer structure (RL + safety filter); Online barrier adaptation | Reasonable computation for iteratively updates control actions; Requires full state feedback; Real-time optimization needed |
| (Wang *et al.*, 2023a) Voltage and frequency control for networked microgrids; Nonlinear ODE system model; Renewable uncertainties and disturbances; State includes voltage, frequency, power | rCLBF; Assumes locally Lipschitz system dynamics and the uncertainty set is convex and bounded | Physics-informed supervised learning; Co-learning of safety certificates and control policy | Local measurements; Real-time capable; Requires state feedback; Validated on load changes and faults |
| (Wan *et al.*, 2023) Load frequency control with high renewable penetration; Parameter uncertainties in inertia, generation loss, and droop gain; Non-stationary operating conditions | Forward invariance via CBF (zeroing and reciprocal variants); Self-tuning safety parameters; Requires feasible compensation action existence | Meta-learning enhanced TD3; Transition post-processing and noise elimination; DRL base + CBF safety layer; GP or uncertainty handling | Real-time QP solving for CBF; GP prediction overhead; Requires partial model knowledge |
| (Shuai *et al.*, 2024) Grid-forming inverter frequency regulation; Nonlinear system dynamics representing Virtual Synchronous Generator; Parameter uncertainties in virtual inertia and damping; Battery storage constraints | Probabilistic Lyapunov stability; Region of Attraction guarantees; GP-based uncertainty quantification; Safe exploration bounds | Model-based RL with Approximate Dynamic Programming; GP dynamics learning; Lyapunov-guided safe exploration within computed ROA | Local measurements only; Robust to parameter variations and external disturbances; Compatible with existing inverter systems |

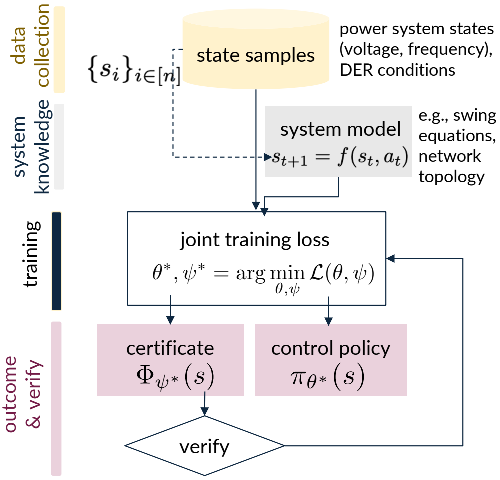


**Figure 3.2:** Overview of the learning-based certificate and joint policy optimization framework. The framework integrates power system domain knowledge/measurements (state samples including voltage, frequency, and DER conditions; system models like swing equations and network topology) with a joint training process. The training minimizes a combined loss function $\mathcal{L}(\theta, \psi)$ that simultaneously learns both a certificate function $\Phi_{\psi^*}(s)$ (e.g., Lyapunov or barrier functions) and a control policy $\pi_{\theta^*}(s)$. The verification stage provides feedback to refine both components, ensuring the learned policy maintains safety constraints while achieving desired performance objectives. This systematic approach enables end-to-end learning of safe and effective control strategies for power system applications.

a basis for characterizing state evolution and enforcing safety conditions. While the generic recipe only requires that the model captures key system characteristics, this model can be enriched by incorporating data-driven refinements, uncertainty quantification, or hybrid representations.

At the core of this stage lies the construction and minimization of a loss function, which is crucial for learning the certificate function.

The key insight in designing the loss function is the conversion of certificate conditions into loss terms. For instance, an inequality condition $a \leq b$ is transformed into a loss term $\max(0, a-b)^2$, while an equality condition $a = b$ becomes $(a-b)^2$. Taking the Lyapunov function as an example, its conditions $\Phi(s_g) = 0$, $\Phi(s) > 0$ for $s \neq s_g$, and $\Delta\Phi(s) < 0$ are translated into loss terms $\Phi(s_g)^2$, $\max(0, -\Phi(s))^2$, and $\max(0, \Delta\Phi(s))^2$, respectively. The overall loss function can be expressed as:

$$\mathcal{L}_{\text{cert}}(\psi) = \Phi_\psi(s_g)^2 + e_1 \sum_i \max(0, -\Phi_\psi(s_i))^2 + e_2 \sum_i \max(0, \Delta\Phi_\psi(s_i))^2$$

where $\psi$ represents the parameters of the neural network encoding the certificate function.

The learning process requires careful consideration of sample coverage, focusing on critical regions such as the boundaries of the safe set for a CBF. Balancing the weights $\{e_1, e_2\}$ of different certificate conditions in the loss function is crucial. Practical enhancements include adaptive sampling, adding robustness margins in certificate conditions, and implementing online policy refinement.

Extending this approach to joint learning of certificates and control policies involves integrating the policy network into the training stage. The loss function is modified to include both certificate and policy components:

$$\mathcal{L}(\psi, \theta) = \mathcal{L}_{\text{cert}}(\psi, \theta) + e_3 \mathcal{L}_{\text{policy}}(\theta)$$

Here, $\psi$ and $\theta$ are the parameters of the certificate and policy networks, respectively. $\mathcal{L}_{\text{cert}}$ enforces certificate conditions using actions generated by the policy $\pi_\theta(s)$, while $\mathcal{L}_{\text{policy}}$ represents the negative expected discounted return: $-\mathbb{E}[\sum_t \gamma^t r(s_t, \pi_\theta(s_t))]$.

This joint optimization allows the certificate and policy to co-evolve, potentially leading to safer and more efficient control strategies. The balance between certificate satisfaction and policy performance can be adjusted through the hyperparameter $e_3$.

### Extensions and Adaptations in Power System Control

The basic approach assumes known dynamics $f(s, a)$, which is often unrealistic. This limitation is evident in the certificate conditions, e.g., $\dot{\Phi}_\psi(s) = \nabla\Phi_\psi(s) \cdot f(s, \pi_\theta(s))$. In practice, $f(s, a)$ is often uncertain or unknown, necessitating approaches that can handle model uncertainty or learn from data directly. This leads to methods that either estimate $f(s, a)$ from data, use robust formulations to account for uncertainty, or learn certificates directly from trajectory data without explicitly modeling the dynamics (Chang *et al.*, 2019).

Recent work has demonstrated how this generic recipe outlined in Fig. 3.2 can be specialized to address the unique challenges encountered in power systems. These studies illustrate how to incorporate physics-informed modeling, leverage known structures such as droop control laws, handle renewable energy uncertainties, and maintain safety and stability under complex operating conditions.

For instance, Zhao *et al.* (2023) focus on transient stability control in systems with synchronous generators and inverter-based distributed generators, ensuring that operational limits on voltage and frequency are strictly maintained. Their approach follows a modified version of the recipe by incorporating an adaptive refinement mechanism for online barrier certificate updates, along with a policy learned via DDPG. Similarly, Wang *et al.* (2023a) target hierarchical control of networked microgrids, blending control Lyapunov and Control Barrier Functions to create a robust and physics-informed loss function that jointly addresses stability and safety. This adaptation leverages well-understood power system dynamics, droop control characteristics, and specific microgrid constraints, showing how the recipe's steps can be enriched by domain knowledge. Moreover, Wang *et al.* (2023a) implement periodic data regeneration and careful sample selection to enhance exploration while respecting operational limits.

Active data collection and hybrid modeling approaches also emerge as key strategies for achieving domain-specific objectives. Shuai *et al.* (2024) consider grid-forming inverter frequency regulation, combining first-principles virtual synchronous generator (VSG) modeling with data-driven Gaussian Process (GP) corrections. They integrate safe exploration into the learning process by focusing on a Region of Attraction (ROA) that guides both data selection and policy refinement. The result is a robust control scheme tailored to realistic parameter uncertainties and storage system constraints, showing how the generic recipe's data collection and model selection steps can be adapted for more complex and uncertain conditions. Similarly, Wan *et al.* (2023) emphasize the interplay between known power system models (e.g., simplified swing equations), GP-based parameter adaptation, and Control Barrier Functions (CBFs) that provide real-time safety layers. Their continuous data acquisition strategy and transition post-processing techniques demonstrate how the recipe's validation and refinement stages can be made domain-aware to handle evolving operational scenarios and variable renewable penetration. Another approach is to combine model-based reinforcement learning with Lyapunov stability theory to handle unknown dynamics in power systems using Gaussian Process (GP) (Shuai *et al.*, 2024).

## 3.3 Practical Considerations

**Structure-Constrained Learning** Structure-constrained learning integrates domain knowledge and constraints into machine learning models. Early demonstrations (Cui *et al.*, 2023) introduce neural architectures enforcing monotonicity via Lyapunov-like constraints. Subsequent work (Feng *et al.*, 2023) adapts these principles to voltage regulation by incorporating voltage-specific constraints such as deadbands, and Liu *et al.* (2024b) combine monotonicity with adversarial training to ensure robustness against malicious inputs. Beyond monotonicity, more general structure-constrained approaches address complex dynamics and uncertainty. For example, Jin and Lavaei (2020) propose gradient-bounded neural networks to ensure well-bounded behavior and incorporate input sparsity aligned with communication topologies under an IQC-based analysis. Similarly, Gu *et al.* (2022) ensure stability for partially observed systems by projecting RNN weights into convex sets derived from IQCs and loop transformations. Such constraints reduce the search space, accelerate stable learning, and improve generalization by embedding inductive biases.

**Computational Efficiency** Computational requirements vary widely with problem scale, device capabilities, and temporal response needs. Some works, such as (Cui *et al.*, 2023), capitalize on structured training methods to achieve several-fold reductions in training times over standard policy gradient baselines. Others, such as (Feng *et al.*, 2023), operate in the millisecond regime for voltage control and rely on relatively simple, localized computations that scale linearly with network size; training may take on the order of minutes for medium-sized systems and longer for large networks, yet still remains manageable. Methods that require solving SDPs for stability certification and weight projections (Jin and Lavaei, 2020; Gu *et al.*, 2022) add overhead offline or online in exchange for safety during exploration, but retain tractable online inference. Approaches such as (Wan *et al.*, 2023) focus on faster convergence via meta-learning techniques, ensuring that training updates can be computed efficiently offline and then rapidly adapted online.

**Scalability** Studies demonstrate scalability from small frequency regulation scenarios to medium and large distribution networks. Decentralized policies and local measurements (Feng *et al.*, 2023; Gu *et al.*, 2022) show near-linear scalability. Other work (Wang *et al.*, 2023a) employ locally controlled DERs with centralized training and distributed execution. Common themes include modularity, local decisions, and distributed architectures to avoid exponential computational growth.

**Generalization** Robustness to variable load profiles, renewable fluctuations, and contingencies is critical. Meta-learning and GP regression (Wan *et al.*, 2023) improve adaptation under diverse conditions. Similarly, (Feng *et al.*, 2023) validate stable performance across scenarios not explicitly trained upon. Zhao *et al.* (2023) provide a generalization bound that with high probability, the learned barrier certificate will generalize to unseen data, under the assumptions of stationary environmen and independent and identically distributed (IID) samples. The trend favors training procedures that systematically promote flexible adaptation to a diverse range of plausible grid states.

**Robustness** Ensuring reliability under adversarial disturbances and uncertainty extends beyond generalization. Robust control techniques such as (Jin and Lavaei, 2020) and (Gu *et al.*, 2022) handle parameter variations, model uncertainty and partial observability. Methods like (Wang *et al.*, 2023a) use robust control Lyapunov barrier functions to handle parameter variations and external disturbances, while (Liu *et al.*, 2024b) address adversarial attacks directly by incorporating adversarial training into the monotone neural architectures. Collectively, these strategies maintain safe, stable operation despite parameter drifts, volatile renewables, and malicious perturbations. These robustness properties are essential for practical deployment of safe MBRL methods in real-world power systems, where resilience against unexpected conditions is paramount for reliable operation.

# 4

# Safe Multi-Agent RL

Recent advances in Multi-Agent Reinforcement Learning (MARL) offer a promising framework for tackling the complexity, scalability, and reliability challenges in modern power systems. Such systems comprise numerous interacting components—distributed energy resources, flexible loads, and control devices—operating under partial observability, non-stationarity, and stringent constraints. MARL naturally fits these environments, coordinating decentralized agents to achieve global efficiency and robustness.

The subsequent sections will delve into key concepts and frameworks in MARL (Sec. 4.1), examine the unique challenges in power systems (Sec. 4.1.2), and discuss techniques and approaches that can enhance coordination, scalability, and, most importantly, safety (Sec. 4.3).

**Notation Primer** Building on previous chapters, we extend to the multi-agent setting for the set of agents $\mathcal{K} = \{1, ..., K\}$. The joint state and action spaces decompose as $\mathcal{S} = \times_{k=1}^{K} \mathcal{S}_k$ and $\mathcal{A} = \times_{k=1}^{K} \mathcal{A}_k$ respectively. For partial observability, each agent $k$ has local observation space $\mathcal{O}_k$ with observation function $O$. Agent-specific rewards $r_k$ and value functions $V_k(\pi_k, \pi_{-k})$ capture individual objectives, where $\pi_k$ is agent $k$'s policy and $\pi_{-k}$ denotes others' joint policy. Safety is enforced through the safe policy set $\Pi_{\text{safe}} = \{\pi : J_c(\pi) \leq \xi\}$, barrier functions $B(s)$, and safety sets $\mathcal{S}_{\text{safe}} = \{s \mid h(s) \geq 0\}$. For power networks, $\mathcal{E}$ represents edges with communication weights $\omega(k, k')$, while $v$, $p^{\text{act}}$, and $p^{\text{react}}$ denote voltage magnitude, active and reactive power. Indices $k$, $j$, $t$, and $i$ consistently refer to agents, buses, timesteps, and iterations respectively.

## 4.1 Perspectives from MARL

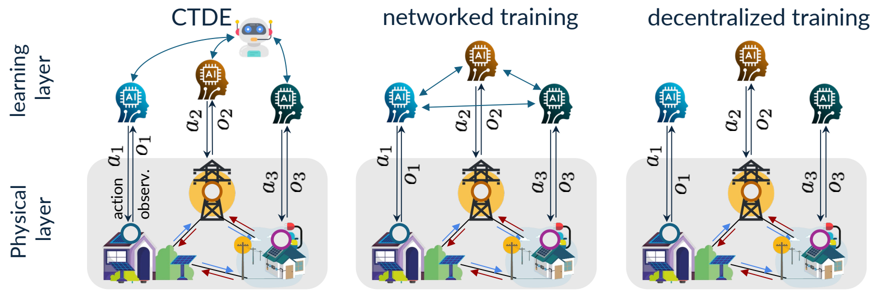
### 4.1.1 MARL Fundamentals

MARL is a framework that extends the single-agent RL paradigm to environments with multiple agents. In MARL, agents interact with the environment and each other, aiming to learn optimal policies that maximize their individual or collective rewards. The Decentralized Partially Observable Markov Decision Process (Dec-POMDP) is a formal model for MARL problems (Oliehoek, Amato, *et al.*, 2016). A Dec-POMDP extends MDP to multi-agent settings with partial observability. Key additions include joint action space $\mathcal{A} = \times_k \mathcal{A}_k$, agent-specific rewards $\{r^k\}_{k\in\mathcal{K}}$, individual observation spaces $\{\mathcal{O}_k\}_{k\in\mathcal{K}}$, and observation probability function $O$.

In a fully cooperative Dec-POMDP, agents share a reward function and seek a joint policy $\pi = \{\pi_k\}_{k\in\mathcal{K}}$ that maximizes the expected discounted cumulative reward (Eq. 1.1). In contrast, competitive or mixed Dec-POMDPs involve agents with individual reward functions $r^k$ aiming to maximize their own expected discounted return $V_k(\pi_k, \pi_{-k}) = \mathbb{E}_{\tau\sim\mathbb{P}(\cdot|\{\pi_k\}_{k=1}^K)}\left[\sum_t \gamma^t r_t^k\right]$ while considering others' policies $\pi_{-k}$, where we use $\tau \sim \mathbb{P}\left(\cdot|\{\pi_k\}_{k=1}^K\right)$ to denote the distribution of trajectory $\tau$ under the joint policy $\{\pi_k\}_{k=1}^K$. The joint policy $\pi^* = \{\pi_k^*\}_{k=1}^K$ represents a Nash equilibrium when $\pi_k^* \in \arg\max_{\pi_k} V_k(\pi_k, \pi_{-k}^*)$.

There are three training regimes in MARL, differed by the flow of information and control among agents.

**Centralized Training with Decentralized Execution (CTDE)**
In CTDE, the training process is centralized, allowing agents to



| | CTDE | networked training | decentralized training |
|---|---|---|---|
| learning speed | high | medium | low |
| comm. require. | high | medium | low |
| scalability | low | medium | high |

**Figure 4.1: Comparison of MARL training paradigms.** The figure distinguishes the physical layer (power grid) from the information/learning layer (agents and their communication patterns). The bottom table summarizes key trade-offs: CTDE (left) relies heavily on global information with high learning efficiency but limited scalability; networked (middle) balances these factors at a medium level; and decentralized (right) uses minimal communication, enabling high scalability, but may slow down learning.

access the global state (e.g., full network states, renewable conditions) and other agents' actions and observations. However, during execution, the agents only have access to their local observations and actions. CTDE enables the learning of coordinated policies while ensuring decentralized execution. This regime has been applied in distribution network voltage regulation (Chen *et al.*, 2021; Shi *et al.*, 2023; Mu *et al.*, 2023), coordinating multiple microgrids for resilience (Qiu *et al.*, 2023), as well as in DSO-VPP operations (Sun and Lu, 2024) and anomalous measurement conditions (Li *et al.*, 2024). This regime aligns well with power systems, where a central entity (e.g., a system operator) can gather comprehensive data offline/online for training, and the distributed agents require only local signals during actual network operation.

**Networked MARL** In networked MARL, agents are connected via a communication network and can exchange information during training and execution. The communication topology can be fixed or dynamic, and agents must learn to communicate effectively to achieve optimal performance. For example, Zhang *et al.* (2024a) consider a fixed communication topology for demand management in distribution networks, where agents learn policies based on local observations while sharing their value function parameters $\theta_k$ with their neighbors for consensus update is: $\theta_k^{i+1} = \sum_{k' \in \mathcal{N}(k)} \theta_{k'}^i$ where $\mathcal{N}(k)$ is the set of neighbors of agent $k$ in the communication graph, and $i$ represent the iteration index. This regime has also been used to achieve efficient multi-microgrid power exchanges (Zhang *et al.*, 2020b), volt-VAR control via distributed devices (Gao *et al.*, 2021), and stable DC microgrid operations (Fan *et al.*, 2023b). Networked MARL suits power systems where peer-to-peer communication among physically adjacent components (e.g., neighboring microgrids, close-by voltage regulators) can enhance coordination, reduce complexity, and scale to larger networks.

**Decentralized MARL** In decentralized MARL, agents (e.g., devices or aggregators) learn independently based on their local observations and rewards. This regime is more scalable and robust to changes in the environment or the number of agents. However, it can suffer from non-stationarity and suboptimal coordination due to the lack of global information. While few works adopt pure decentralized MARL (with no communication or central training), some resemble a more decentralized approach by limiting global information and relying heavily on local measurements and local interactions (Zhang *et al.*, 2020b; Fan *et al.*, 2023b).

Figure 4.1 illustrates these three training paradigms and their key trade-offs in terms of learning speed, communication requirements, and scalability. In addition to how information and responsibilities are distributed among agents, another important dimension is whether learning proceeds on-policy or off-policy, which affects sample efficiency, stability, and adaptability under changing power system conditions.

**Off-Policy Methods** Off-policy method has been widely used, often in the settings of CTDE (e.g., MADDPG or TD3) (Shi *et al.*, 2023; Qiu *et al.*, 2023; Mu *et al.*, 2023; Sun and Lu, 2024; Li *et al.*, 2024) or networked MARL (Zhang *et al.*, 2020b; Gao *et al.*, 2021; Fan

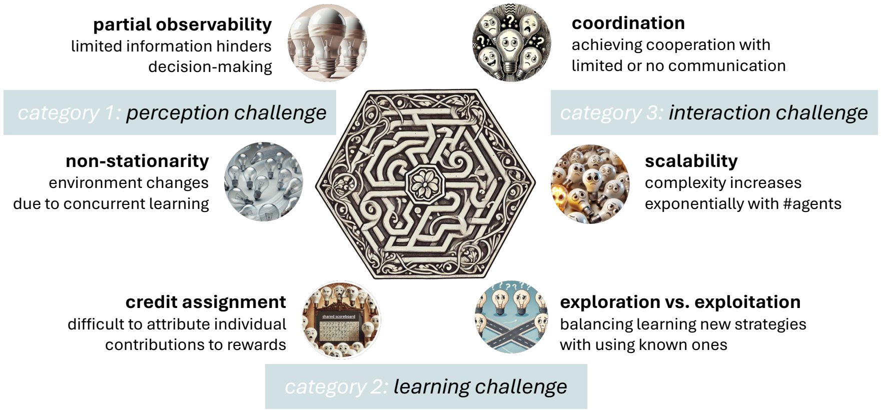

partial observability
limited information hinders decision-making
coordination
achieving cooperation with limited or no communication
category 1: perception challenge
category 3: interaction challenge
non-stationarity
environment changes due to concurrent learning
scalability
complexity increases exponentially with #agents
credit assignment
difficult to attribute individual contributions to rewards
exploration vs. exploitation
balancing learning new strategies with using known ones
category 2: learning challenge

**Figure 4.2: MARL Maze.** Illustration of key MARL challenges, emphasizing the complexity and interconnectedness of these issues.

*et al.*, 2023b). By using experience replay buffers, these methods can learn from past trajectories and revisit diverse operational scenarios (e.g., varying load patterns, renewable uncertainties, and evolving market prices), improving sample efficiency.

**On-Policy Methods** On-policy learning (e.g., policy gradient based variants), while less common, can align closely with real-time operational constraints, as policies are updated directly from the trajectories they generate. This may offer stable adaptation to current conditions but can be less sample-efficient. Some examples include (Chen *et al.*, 2021; Zhang *et al.*, 2024a), where agents learn continuously from their current policy's behavior.

The choice of training regime often reflects the complexity and scale of the power system application, the availability of global versus local information, and the balance sought between achieving high performance, ensuring operational safety, and maintaining scalability.

### 4.1.2 Key Challenges in Power Systems

MARL faces three primary categories of challenges (illustrated in Figure 4.2):

**Perception Challenges** Perception challenges encompass partial observability and non-stationarity. In partially observable environments, agents receive observations $o_k \sim O(s, a_k)$ (e.g., its own demand, local generation, and carbon emissions) instead of the full state $s$ (e.g., other agents' actions, network congestion, or global power flow), leading to hidden state information. This is often modeled as a Dec-POMDP, where agents must make decisions based on local observations. Non-stationarity is inherent due to continuously changing loads, intermittent renewable generation, and time-varying market signals. It can also arise as other agents' policies evolve during learning, violating the Markov assumption and making the transition function time-dependent (i.e., $\mathbb{P}(s'|s, a, t)$ depends on $t$). These conditions make stable policy learning more difficult, as agents must adapt to evolving environments and manage uncertainty in both observation and system dynamics.

**Learning Challenges** Learning challenges include credit assignment and the exploration-exploitation dilemma. Attributing system-level improvements (e.g., stable voltages, reduced losses, or lower carbon emissions) to individual agent actions is complex in interconnected power networks. This credit assignment problem stems from the intertwined effects of multiple devices and control strategies. Moreover, exploration in a critical infrastructure poses a significant risk: naive experiments might lead to voltage/frequency excursions and compromise safety or compliance with grid codes. Balancing the need for exploration with the imperative for stable operations (exploitation) presents a delicate challenge.

**Scalability and Coordination Challenges** Scalability and coordination challenges intensify as the number of agents increases. The joint action space grows exponentially: $|A| = |A_1| \times |A_2| \times \ldots \times |A_K|$, making centralized approaches intractable for large systems. Large-scale networks must integrate strategies that allow local decision-making to aggregate into stable system-wide outcomes. Achieving such coordinated control across a large number of agents, each with limited observability and potentially conflicting objectives, is a key hurdle—one that highlights the importance of scalable and communication-aware MARL frameworks in the evolving landscape of modern power grids.

These challenges are connected, with methods often addressing multiple issues simultaneously.

## 4.2 Techniques for Addressing MARL Challenges

### 4.2.1 Tackling Perception Challenges

To overcome *partial observability*, several approaches emerge.

Recurrent networks, such as Gated Recurrent Units (GRUs) (e.g., (Mu *et al.*, 2023; Chen *et al.*, 2022; Shi *et al.*, 2023)) and Long Short-Term Memory (LSTM) (e.g., (Chen *et al.*, 2021)), maintain an internal state to integrate temporal modeling and can effectively encode history to extract relevant features.

Information sharing among neighbors, such as the use of networked MARL and consensus approaches, can help broaden agent's view beyond purely local measurements (Zhang *et al.*, 2020b; Gao *et al.*, 2021; Fan *et al.*, 2023b; Zhang *et al.*, 2024a). Graph-based models such as GCN can also incorporate topological information when integrating neighbor data (Mu *et al.*, 2023). Communication methods such as CommNet (Sukhbaatar, Fergus, *et al.*, 2016) aim to learn a communication protocol to enable information sharing, which has been adapted by (Chen *et al.*, 2021) to foster the collaborations among neighboring agents. Learning a surrogate model using Sparse Variational Gaussian Processes (SVGP) to create a simulation environment for MARL (Li *et al.*, 2023) can reduce real-world communication and data collection.

To handle *nonstationarity*, Hernandez-Leal *et al.* (2017) discuss five categories of approaches, among which ignoring (assuming stationarity) and forgetting (updating based on recent observations), and a few works on responding/learning opponent models are most common in power systems.

Attention mechanisms, exemplified by Multi-Actor-Attention-Critic (MAAC) (Iqbal and Sha, 2019), focus on relevant parts of observations

**Table 4.1:** Comparison of Multi-Agent Control Approaches in Power Systems

| **Problem Setup** | **MARL Algorithm** | **Enhancements** | **Comments** |
|---|---|---|---|
| (Chen *et al.*, 2021): Microgrid secondary voltage regulation; Cooperative agents with local measurements; Voltage bounds | On-policy actor-critic; CTDE; Action smoothing; Experience replay | Spatial discount factor for scalability; Differentiable comm.; LSTM for partial observability | Tested under load variations; 35.8ms inference time for 40 agents |
| (Shi *et al.*, 2023): Distribution networks (33–322 bus); Voltage control via PV inverters; Cooperative agents with partial observability; Voltage constraints | MADDPG-based; Safety layer with action correction; Off-policy CTDE; Experience replay | Parameter sharing for scalability; GRU for partial observability; Centralized critic | Tested across seasonal variations; Quick safety computations |
| (Zhang *et al.*, 2024a): Low-carbon demand management; Carbon emission constraints; Bi-level control (aggregate load agents, distribution network operator) | Actor-critic; Trust region updates for constraints; On-policy training | Consensus-based coordination; Fixed communication topology; Network-based value updates | Privacy preservation; Handles renewable uncertainty; Carbon reduction metrics |
| (Zhang *et al.*, 2020b): Power management optimization; Cooperative agents with local observations; Voltage, current, and operational constraints | Gaussian policy functions; Gradient-based safe policy learning; Off-policy distributed training | Distributed consensus-based training; Backtracking for constraint satisfaction | Privacy preservation between MGs; Real-time decision making (1.4s vs 145.5s central opt.) |
| (Gao *et al.*, 2021): Volt-VAR control via voltage regulators, capacitors, and tap changers; Cooperative agents with local observations; Operational constraints | Maximum entropy RL; Off-policy training with randomized consensus protocol; Experience replay | Comm.-efficient consensus strategy; Local reward decomposition with value function consensus | Robust against agent/communication failures |
| (Qiu *et al.*, 2023): Resilience-oriented coordination of networked microgrids; Cooperative agents with Dec-POMDP; Voltage and power flow constraints | Shapley Q-value + DDPG; Off-policy CTDE; Gaussian noise exploration; Experience replay | Shapley value for credit assignment; Centralized critic; Power exchange coordination | Real-time deployment (≈0.6s); Resilience index metric; Handles uncertainties |
| (Fan *et al.*, 2023b): DC microgrid OPF with multiple DGs; Cooperative agents with neighbor-based partial observability; Power and voltage operational bounds | TD3-based distributed architecture; Off-policy training with experience buffer initialization | Neighbor-based communication topology; Safe exploration through PI controller initialization | Handles topology changes and dynamic loads |
| (Mu *et al.*, 2023): Distribution network voltage control via PV inverters; Cooperative agents with Dec-POMDP; Voltage constraints | MADDPG-based; Barrier function for safety; Off-policy CTDE with parameter sharing | GCN for topology embedding; GRU for temporal dependencies | Tested across topologies; Metrics include controllable ratio and power loss |
| (Sun and Lu, 2024): Dual-layer Stackelberg game between DSO and VPPs; Mixed cooperative-competitive agents with privacy requirements; AC power flow and voltage constraints | Parameter-sharing TD3; Off-policy CTDE; Prioritized experience replay | Privacy-preserving multi-agent joint Q-value function; Improved convergence through shared experiences | Demonstrated real-time capability; Privacy preservation in market operations |
| (Li *et al.*, 2024): Voltage control and economic dispatch under anomalous measurements; Cooperative agents with local observations; Operational constraints | MAAC-based; Off-policy CTDE framework; Experience replay | Confederate image technology; Multi-head graph attention for topology; GRU for trajectory history | Robust to measurement anomalies; Metrics include economic cost, voltage deviation |

or other agents' information, aiding in addressing the perception challenges. Belief state methods explicitly handle uncertainty by maintaining probability distributions over possible states.

Opponent modeling, as seen in DRON (Deep Reinforcement Opponent Network) (He *et al.*, 2016) and MADDPG (Lowe *et al.*, 2017), addresses non-stationarity by predicting other agents' behaviors. Li *et al.* (2024) develop a technique namedconfederate image technology to maintain a model of other agents.

CTDE allows agents to share information during training but act based on local observations during execution. MADDPG (Multi-Agent Deep Deterministic Policy Gradient) (Lowe *et al.*, 2017), a popular CTDE method, uses a centralized critic conditioned on all agents' observations and actions, while the actor only accesses local information. Although primarily applied to cooperative settings (e.g., (Mu *et al.*, 2023)), MADDPG can also handle mixed cooperative-competitive environments. Off-policy learning enhances stability by learning from past experiences. Examples include MASAC (Multi-Agent Soft Actor-Critic) (Chen *et al.*, 2024), off-policy maximum entropy RL (Gao *et al.*, 2021), and Twin TD3 (Delayed Deep Deterministic Policy Gradient) (Fan *et al.*, 2023b; Sun and Lu, 2024).

These methods often work synergistically, providing benefits across multiple MARL challenges.

### 4.2.2 Credit Assignment and Exploration-Exploitation

For credit assignment, MARL employs several key approaches. Difference Rewards evaluate an agent's contribution by comparing the global reward with and without the agent's action. Counterfactual Multi-Agent Policy Gradients, exemplified by COMA, use a centralized critic to compute a counterfactual baseline for each agent. Shapley Value Methods, such as SQDDPG (Wang *et al.*, 2020), incorporate game-theoretic concepts to fairly attribute contributions to each agent. Qiu *et al.* (2023) employ Shapley Q-values to more explicitly measure each agent's marginal contribution, directly addressing credit assignment for power system resilience and microgrid coordination.

To address exploration-exploitation dilemma, maximum entropy frameworks, as seen in (Gao *et al.*, 2021; Li *et al.*, 2023), can balance exploration with exploitation by adjusting the weight for the entropy term. Gaussian noise perturbations to the actor's output, along with large replay buffers, have been used (Qiu *et al.*, 2023; Fan *et al.*, 2023b). Off-policy learning with replay buffer such as those used in MATD3 (Ackermann *et al.*, 2019) algorithm can also help (Chen *et al.*, 2022).

### 4.2.3 Scalability and Coordination

To enhance scalability, parameter sharing is a common approach, where agents share network parameters for value function or policy estimation (Mu *et al.*, 2023; Sun and Lu, 2024). This allows leveraging data from all agents to update a single shared network, improving scalability and reducing policy oscillations. Combining parameter sharing with Graph Convolutional Networks (GCNs) can further incorporate topology information (Mu *et al.*, 2023). Spatial discount factors (Chen *et al.*, 2021) encourage agents to consider the impact of their actions on neighboring agents, limiting the state/action space span.

To tackle the coordination challenge, (Zhang *et al.*, 2020b; Gao *et al.*, 2021; Zhang *et al.*, 2024a; Mu *et al.*, 2023) rely on decomposing the global control task into local decisions informed by neighbor-to-neighbor communication or CTDE. By operating on a graph structure where agents represent network nodes, these works ensure that complexity grows linearly or sub-linearly with system size. (Chen *et al.*, 2021; Mu *et al.*, 2023) implement neighbor-to-neighbor communication, where agents exchange partial information to coordinate voltage references.

Incentive mechanisms can be designed to encourage collaboration. A cooperative bi-level framework, introducing an asymmetric Markov game to align agent objectives and guide equilibrium behaviors, along with a bi-level actor-critic algorithm for real-time control, is proposed in (Hong *et al.*, 2024). Similarly, Sun and Lu (2024) adopt a bi-level approach to balance operational safety and market participants' interests. While existing approaches use penalty functions and global reward signals to promote cooperation and align objectives, Lin *et al.* (2024) introduce Markov Signaling Game, a framework for studying strategic incentive-compatible communication between a sender and a receiver.

The signaling gradient and extended obedience constraints help learn efficient and stable policies under information asymmetry.

### 4.2.4 Communication Efficiency and Robustness

Communication allows agents to share information and coordinate actions, but it must be done efficiently. While decentralized training such as (Omidshafiei *et al.*, 2017) is available, selective communication is common, where agents only communicate a subset of relevant information, such as value/policy data (Gao *et al.*, 2021) or encoded state information (Chen *et al.*, 2021). In structured communication, such as networked MARL (e.g., (Gao *et al.*, 2021; Fan *et al.*, 2023b)), each agent only needs to communicate with its neighbors. Agents can also learn communication protocols end-to-end, such as using differentiable communication (Chen *et al.*, 2021).

To handle agent and communication failures, Gao *et al.* (2021) propose constructing replacement states using historical averages and the agent's own policy networks to maintain operations. The impact of communication topology changes on learning performance is studied in (Fan *et al.*, 2023b).

### 4.2.5 Discussion of MARL Methods

Key insights from MARL research highlight that many methods address multiple challenges simultaneously, and the choice of method often depends on specific problem characteristics. CTDE, as seen in MADDPG and COMA (Foerster *et al.*, 2018), addresses non-stationarity, credit assignment, and scalable coordination by leveraging global information during training while allowing for decentralized execution.

Attention Mechanisms, exemplified by MAAC and ATOC (Jiang and Lu, 2018), help with partial observability, credit assignment, and scalability by selectively focusing on relevant information. Value Decomposition methods like VDN (Sunehag *et al.*, 2018) and QMIX (Rashid *et al.*, 2020) focus on credit assignment and scalability by decomposing team value functions into individual components. Communication-Based Methods, such as DIAL (Foerster *et al.*, 2016), aid in overcoming partial observability, coordinating exploration, and enabling scalable information sharing through learned communication protocols.

Combining multiple approaches frequently yields the best results, as different methods can complement each other's strengths. However, trade-offs exist between scalability, adaptability, and computational complexity, requiring careful consideration in method selection. Real-world applications often necessitate considering safety, adding another layer of complexity to method choice, which we will address next.

## 4.3 Safety Considerations in MARL

Safe Multi-Agent Reinforcement Learning (MARL) enhances traditional MARL by incorporating safety constraints through cost functions $c$ and their associated thresholds $\xi$. These safety constraints can be implemented from either global or local perspectives, each serving distinct safety requirements in multi-agent systems.

From a global perspective, safety constraints can be imposed on the global state $s = (s_1, ..., s_K)$ or joint action $a = (a_1, ..., a_K)$ of the multi-agent system. This ensures system-wide safety properties are maintained. For example, Zhang *et al.* (2024a) demonstrate this through system-wide carbon emission constraints on the global state.

Local safety focuses on constraints specific to each agent's local state $s_k$ and action $a_k$. In power systems, local state constraints frequently appear as frequency deviation limits in microgrids (Xia *et al.*, 2022; Liu *et al.*, 2024b), voltage bounds (Shi *et al.*, 2023; Zhang *et al.*, 2024a), and operational constraints for energy storage systems and distributed generators (Xia *et al.*, 2022; Shi *et al.*, 2023). Edge-based safety constraints, such as branch flow limits, are not strictly local but govern interactions between neighboring agents' states and actions (Zhang *et al.*, 2020b).

Table 4.2 provides a survey of safety constraints, as well as safe MARL methods to be discussed next.

**Table 4.2:** Comparison of Safety Approaches in MARL for Power Systems

| **Safety Constraints** | **Safe MARL Method** | **Implementation Details** |
|---|---|---|
| (Zhang *et al.*, 2020b): Voltage limits, branch current flows, DG/ESS operational bounds. Local constraints. | Distributed consensus-based optimization with gradient-based safe learning. Backtracking mechanism for constraint satisfaction. | Real-time decision making (1.4s) on 33-bus distribution network. Requires neighbor communication for constraint coordination. |
| (Gao *et al.*, 2021): Voltage bounds, device switching limits, neighbor power loss constraints. Local operational constraints. | Maximum entropy reinforcement learning with consensus-based coordination. Penalty-based reward structure for safety enforcement. | Robust against agent failures while maintaining safety. Communication-efficient consensus. |
| (Xia *et al.*, 2022): Frequency bounds, local ESS operational bounds (SOC, SOH, power limits). | Dual-network safety scheme (evaluation + guidance networks). Predictive violation detection with safe action generation. | Decentralized execution (0.01s computation time). Handles renewable uncertainty while maintaining frequency stability. |
| (Fan *et al.*, 2023b): Power and voltage operational bounds in DC microgrids. Neighbor-based safety coordination for multiple distributed generators. | Safe exploration through PI controller initialization. Distributed architecture with neighbor communication for safety coordination. | Effectively handles dynamic loads and topology changes while maintaining safety constraints. Requires initial safe exploration data from PI controllers. |
| (Shi *et al.*, 2023): Local voltage constraints with PV inverter reactive power limits. Coordinated voltage control across network. | Safety layer with quadratic programming for action corrections. Centralized safety coordination with first-order approximation. | Scalability from 33-bus to 322-bus systems. Quick safety computations. Tested across seasonal variations. |
| (Zhang *et al.*, 2024a): Global carbon emission limit combined with local voltage constraints. Bi-level control structure. | Trust region policy optimization with constrained updates. Consensus-based coordination for safety enforcement. | IEEE 33-bus and 123-bus networks. Handles renewable uncertainty while maintaining safety constraints. |

### 4.3.1 Constrained Optimization Approaches

#### Lagrangian Methods for Safe MARL

We introduce a general framework for safe MARL, leveraging general utilities (see, e.g., (Ying *et al.*, 2024)):

$$\begin{aligned} \max_{\theta} \quad & F(\theta) := \frac{1}{K}\sum_{k=1}^{K} f_k(d_k^{\pi_\theta}) \\ \text{s.t.} \quad & G_k(\theta) := g_k(d_k^{\pi_\theta}) \geq 0, \quad \forall k \in [K] \end{aligned}$$

where $\theta = \{\theta_k\}_{k\in[K]}$ denotes the joint policy parameters. Here, $f_k$ and $g_k$ are general utilities (i.e., nonlinear functions) of the local state-action occupancy measure $d_k^{\pi_\theta}$ for agent $k$, which is defined as:

$$d_k^{\pi_\theta}(s_k, a_k) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_k^t = s_k, a_k^t = a_k | \pi_\theta, s^0 \sim \rho).$$

The general utilities include the standard cumulative reward/cost formulation when choosing $f_k = \langle d_k^{\pi_\theta}, r^k \rangle$, where $r^k$ is the vector of reward for each state-action pair for agent $k$, but also include other settings such as imitation learning and exploration.

To solve this constrained optimization problem, primal-dual optimization techniques use the Lagrangian function (Zhang *et al.*, 2020b; Gu *et al.*, 2023; Ying *et al.*, 2024):

$$L(\theta, \lambda) = F(\theta) + \frac{1}{K} \sum_{k \in \mathcal{K}} \lambda_k G_k(\theta). \tag{4.1}$$

where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_K)$ are the Lagrange multipliers associated with the local constraints. The algorithms alternate between updating the policy parameters $\theta$ (primal update) and the Lagrange multipliers $\lambda$ (dual update). For the primal update, Gu *et al.* (2023) propose the Multi-Agent Constrained Policy Optimization (MACPO) algorithm, which solves a constrained optimization problem for each agent $k$ using trust region methods. Zhang *et al.* (2020b) introduce distributed consensus-based algorithm using networked exchange of dual variables that allows for efficient handling of edge-based constraints without requiring a fully centralized approach.

A key insight that enables efficient implementation is the spatial correlation decay property in many multi-agent systems. This property recognizes that an agent's influence on others typically diminishes with distance—much like how changes in one part of a power grid have decreasing effects on distant areas. Ying *et al.* (2024) leverage this natural decay in influence and develop Scalable Primal-Dual Actor-Critic method for the primal update. Here, each agent $k$ can effectively estimate its policy gradient using information from only its $\kappa$-hop neighborhood:

$$\hat{\nabla}_{\theta_k} L(\theta, \lambda) = \mathbb{E}\left[ \sum_{k' \in \mathcal{N}^\kappa(k)} \nabla_{\theta_k} \log \pi_{\theta_k}(a_k|s) \cdot (\hat{Q}_{f_{k'}}^{\pi_\theta}(s, a) + \lambda_{k'} \hat{Q}_{g_{k'}}^{\pi_\theta}(s, a)) \right],$$

where $\mathcal{N}^{\kappa}(k)$ denotes the $\kappa$-hop neighborhood of agent $k$, and $\hat{Q}^{\pi_\theta}_{f_{k'}}$ and $\hat{Q}^{\pi_\theta}_{g_{k'}}$ are truncated shadow Q-functions. This localized computation efficiently approximating the global quantities using local information within a $\kappa$-hop neighborhood.

For the dual update, a common approach is to use projected gradient descent:

$$\lambda^{i+1} = [\lambda^i - \eta_\lambda \nabla_\lambda L(\theta^i, \lambda^i)]_+, \tag{4.2}$$

where $\eta_\lambda$ is the step size. This can be done in a centralized fashion where all the dual variables $\lambda$ are updated jointly (Gu *et al.*, 2023), or in a decentralized way where each agent $k$ updates its own Lagrange multiplier $\lambda_k$ based on its local constraint information by leveraging the spatial correlation decay property (Ying *et al.*, 2024).

### 4.3.2 Trust-Region Method

Adapting trust region methods to MARL introduces several challenges, including decentralized decision-making, partial observability, non-stationarity, coordination, and scalability.

Zhang *et al.* (2024a) develop Consensus Multi-Agent Constrained Policy Optimization (CMACPO) to address these challenges. In this framework, each agent $k$ maintains a local policy $\pi_k^i$ at iteration $i$ and makes decisions based on local observations. For each agent, the method uses the single-agent CPO formulation (see (2.1), where $\pi_k^i$ is used for agent $k$'s policy at iteration $i$), followed by the same linearization approach as single-agent CPO. To address partial observability and promote coordination, CMACPO implements a networked communication structure where agents share information with neighbors. A key innovation is the consensus mechanism that aligns value function estimates across agents:

$$\phi_k^{i+1} = \sum_{k' \in \mathcal{N}(k)} \omega(k, k') \phi_{k'}^i$$

where $\phi_i^k$ represents the local value function parameters and $\omega(k, k')$ denotes the communication weight between agents $k$ and $k'$. This consensus step helps mitigate non-stationarity by promoting consistent value estimation across the network while maintaining the decentralized nature of the algorithm. This method is applied to optimize the behavior of aggregated flexible loads (agents) in a distribution network. The trust region approach ensures stable learning, the safety constraints enforce carbon emission limits, and the consensus mechanism allows for coordination among loads while respecting the network structure.

Shi *et al.* (2023) combine trust region concepts with safety constraints in a MARL setting, using a centralized safety layer during training to guide decentralized policies towards safe behavior. The trust region idea is used in the policy update mechanism:

$$\max_{\theta_k} \mathbb{E}_{s\sim d^{\pi}}[Q^{\pi_k}(s,a) - \mu D_{KL}[\mathcal{N}(a_k,\sigma_a^2)\|\mathcal{N}(a_k^{\text{safe}},\sigma_{\text{safe}}^2)]] \tag{4.3}$$

where $Q^{\pi_k}$ is the action-value function under policy $\pi_k$, $\mathcal{N}(a_k,\sigma_a^2)$ and $\mathcal{N}(a_k^{\text{safe}},\sigma_{\text{safe}}^2)$ are the Gaussian distributions of actions from current and computed safe policies with corresponding means and variances, respectively, and $\mu$ is the penalty coefficient for the KL-divergence term. The key difference with standard CPO is that this formulation constrains the policy to stay close to computed safe actions rather than the previous policy iterate. This allows for safe, decentralized execution in voltage control tasks, addressing the challenges of maintaining voltage stability in complex power distribution networks.

### 4.3.3 Control-Theoretic Approaches

Distributed and decentralized CBFs have emerged as a promising approach for ensuring safety in MARL systems. By leveraging the power of GNNs, these methods can effectively handle large-scale, dynamic environments while maintaining scalability and generalizability (Qin *et al.*, 2021; Zhang *et al.*, 2023e).

For a multi-agent system with $K$ agents, consider the joint state space $\mathcal{S} = \times_{k=1}^{K}\mathcal{S}_k$ and individual state-observation spaces $\mathcal{X}_k := \mathcal{S}_k \times \mathcal{O}_k$, where each agent $k$ has state $s_k \in \mathcal{S}_k$, local observation $o_k \in \mathcal{O}_k$. Let the dynamics be $\dot{s}_k = f_k(s_k, a_k)$. We define the safe set as $\mathcal{X}_k^{\text{safe}} = \{(s_k, o_k)|h(s_k, o_k) \geq 0\}$, and dangerous set as $\mathcal{X}_k^{\text{unsafe}} = \{(s_k, o_k)|h(s_k, o_k) < 0\}$. The key idea behind decentralized CBFs is to assign a local CBF $B_k$ to each agent $k$ in the system. These local CBFs capture the safety constraints between the agent and its neighbors, allowing for a distributed safety framework.

**Definition 4.1** (Decentralized Control Barrier Function)**.** A continuously differentiable function $B_k : \mathcal{X}_k \to \mathbb{R}$ is a decentralized control barrier function if it satisfies:

1) $\forall (s_k, o_k) \in \mathcal{X}_k^{\text{safe}}$, $B_k(s_k, o_k) \geq 0$;
2) $\forall (s_k, o_k) \in \mathcal{X}_k^{\text{unsafe}}$, $B_k(s_k, o_k) < 0$;
3) $\forall (s_k, o_k) \in \{(s_k, o_k) | B_k(s_k, o_k) \geq 0\}$, implies that:

$$\sup_{a_k} \nabla_{s_k} B_k(s_k, o_k)^\top f_k(s_k, a_k) + \nabla_{o_k} B_k(s_k, o_k)^\top \dot{o}_k + \alpha(B_k(s_k, a_k)) \geq 0,$$

where $\alpha$ is a class $\mathcal{K}$ function.[1]

The fundamental distinction from traditional CBFs (c.f., Theorem 3.2) lies in the decentralized architecture. While a traditional CBF operates on the full state space $\mathcal{S}$ with complexity growing exponentially in $K$, a decentralized CBF maintains constant complexity by operating only on local information $\mathcal{X}_k$. This enables scalable implementation without requiring centralized coordination. Global safety emerges from local contracts: when each agent satisfies its local CBF condition, the composition of these local guarantees ensures system-wide safety.

To handle variable numbers of neighbors and enable efficient learning, graph neural networks are employed to parameterize the local CBFs $B_k$ and the control policies $\pi_k$. The GNNs take the agent's state and observations of its neighbors as input, and output the CBF value and control action. The use of GNNs allows for permutation invariance and scalability to large-scale systems (Qin *et al.*, 2021; Zhang *et al.*, 2023e).

The joint learning of the CBFs and control policies is typically formulated as a constrained optimization problem, where the objective is to minimize the violation of the local CBF conditions while ensuring task performance summed over all agents. Using CTDE, this closely resembles the single-agent method discussed in Sec. 3.2. By decomposing

[1] Here, $\dot{o}_k$ is the time derivative of the observation, which represents coupling between agents. While there is no explicit expression for this term, it can be approximated in the learning process, e.g., $\dot{o}_k \approx \frac{o_k(t+\delta_{\text{disc}}) - o_k(t)}{\delta_{\text{disc}}}$.

the safety constraints into local CBFs, the approach can handle large-scale multi-agent systems with reduced computational complexity.[2] The learned CBFs and policies can generalize well to unseen scenarios, such as different numbers of agents or new environments.

However, there are also practical considerations to keep in mind. First, while the framework is decentralized, agents still need to exchange information with their neighbors to evaluate the local CBFs and compute control actions. Also, decentralized CBFs may lead to conservative behavior, as each agent considers its safety independently without explicit coordination.

### 4.3.4 Distributed Optimization Perspective

Distributed optimization techniques can be effectively combined with CBFs to achieve safe distributed control in multi-agent systems (Tan and Dimarogonas, 2021; Bai *et al.*, 2024). Recent advancements in distributed CBFs and distributed optimization reveal a common principle: maintaining feasibility of all constraints (both local and global) at every iteration (not just at convergence), also known as *distributed feasible primal methods*. This is crucial for safety-critical systems where intermediate infeasible solutions could lead to system failures.

This is exemplified in the distributed CBF approach by (Tan and Dimarogonas, 2021) and the Distributed Feasible Method (DFM) (Wu *et al.*, 2023) for optimization. Both approaches reformulate centralized problems into distributed forms. The distributed CBF transforms the centralized QP:

$$\min_a \sum_{k \in \mathcal{K}} \frac{1}{2} \|a_k - a_k^{\text{nom}}\|^2 \quad \text{s.t.} \quad \sum_{k \in \mathcal{K}} A_k^\top a_k + \sum_{k \in \mathcal{K}} b_k \leq 0$$

[2] In partially observable environment, safety is ensured under some key assumptions (e.g., (Qin *et al.*, 2021)): **1)** Local and reciprocal observability: Each agent $k$ can observe nearby agents within a radius significantly larger than the safety threshold. If agent $k$ can observe agent $k'$, then $k'$ can also observe $k$ and **2)** Transitive Safety: Safety between directly unobservable agents is ensured through the chain of pairwise safe interactions.

Into a distributed form with auxiliary variable $y$:

$$\begin{aligned} &\min_{a,y} \sum_{k \in \mathcal{K}} \frac{1}{2} \|a_k - a_k^{\text{nom}}\|^2 \\ &\text{s.t.} \quad A_k^\top a_k + \sum_{k' \in \mathcal{N}(k)} (y_k - y_{k'}) + b_k \leq 0, \quad \forall k \in \mathcal{K} \end{aligned}$$

where $a_k^{\text{nom}}$ is the nominal (desired) control action for agent $k$, $A_k$ and $b_k$ are the CBF constraint matrix and offset for agent $k$, respectively,[3] and $y_k$ is the auxiliary variable.

This allows decomposition of the coupled problems into local subproblems using distributed optimization techniques, where each agent maintains its copy of auxiliary variables and only need to communicate with its neighbors. The global feasibility is maintained because:

$$\sum_{k \in \mathcal{K}} A_k^\top a_k + \sum_{k' \in \mathcal{N}(k)} (y_k - y_{k'}) + b_k = \sum_{k \in \mathcal{K}} A_k^\top a_k + \sum_{k \in \mathcal{K}} b_k \leq 0$$

DFM, on the other hand, uses auxiliary variables $q_{kk'}$ along each edge that satisfies $\sum_{k' \in \mathcal{N}(k)} A_{k'} q_{kk'} = 0$ to maintain feasibility during updates.

A key focus is to ensure constraint satisfaction throughout the optimization process. For instance, Bai *et al.* (2024) develop a distributed hybrid gradient projection ADMM algorithm and prove its convergence to the optimal solution under appropriate conditions while guaranteeing coupling constraint satisfaction at each iteration. Tan and Dimarogonas (2021) propose a normalized subgradient flow to update the auxiliary variables, which converges to the optimal solution in finite time while maintaining constraint satisfaction. Note that the finite time convergence is a property not typically found in standard distributed

[3] The terms $A_k$ and $b_k$ in the distributed CBF formulation emerge from the safety set defined by a continuously differentiable barrier function $B(s) \geq 0$. For a control affine system $\dot{s} = f_s(s) + f_a(s)a$, the CBF condition requires $\nabla_s B(s)^\top (f_s(s) + f_a(s)a) + \alpha(B(s)) \geq 0$, which can be rearranged into the linear constraint form $A_k^\top a_k + b_k \leq 0$. Here, $A_k^\top = -\nabla_s B(s)^\top f_a(s)$ captures the gradient of $B$ with respect to control input, and $b_k = -\nabla_s B(s)^\top f_s(s) - \alpha(B(s))$ encompasses the drift terms and class $\mathcal{K}$ function effects. Crucially, these terms must be locally computable by each agent using only information from its neighbors, as assumed in the distributed implementation.

optimization algorithms which often have asymptotic convergence. Tan *et al.* (2024) extend (Tan and Dimarogonas, 2021) to a more general class of distributed optimization problems, not limited to CBF-QPs and handles multiple coupling constraints. It also introduces a continuous-time algorithm and uses a subgradient-based approach for updating auxiliary variables, which can handle a broader class of non-smooth objective functions. These approaches provide a powerful tool for achieving safe distributed control in multi-agent systems and have promising applications in safe MARL.

# 5

# Safety System Architecture

The pursuit of extreme safety in power systems presents significant challenges when applied to RL systems. Runtime Assurance (RTA) architectures have emerged as a crucial framework, providing an implementable safety system architecture that complements broader safe and multi-agent RL approaches (Schierman *et al.*, 2015; Hobbs *et al.*, 2023). These architectures provide formal guarantees by monitoring and modifying control actions in real-time, effectively bridging the gap between high-performance learning controllers and safety-critical power system operations.

In this chapter, we introduce the two main categories of RTA architectures—Simplex and Safety Filter—and examines their implementations in power systems (Sec. 5.1 ). Sec. 5.2 discusses various approaches for integrating safety mechanisms into the learning process itself. Finally, Sec. 5.3 addresses practical considerations including innocuity, viability, and nuisance-freedom, while highlighting future research directions in this domain.

## 5.1 RTA Architectures for Safe Learning Systems

RTA architectures for safe learning systems primarily fall into two categories: Simplex Architecture and Safety Filter (SF) (see Fig. 5.1 for an illustration). In essence, Simplex provides a “fallback” mechanism (switching between controllers), while SF provides a “correction” mechanism (continuously adjusting actions). While Simplex is particularly suitable for applications demanding robust safety guarantees and formal verification, SF is potentially less conservative as it can make minimal modifications rather than completely switching to a backup controller. The following sections detail these architectures and their implementations in power system applications (see Table 5.1 for a comparison).

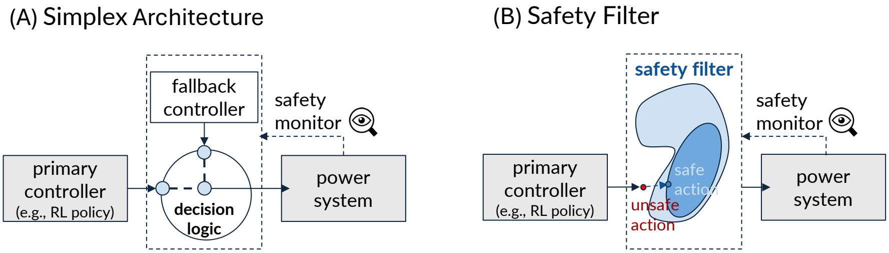


**Figure 5.1: Comparison of RTA architectures.** (A) Simplex Architecture employs a switching mechanism between a primary learning controller and a backup controller based on a safety monitor’s decision logic. (B) Safety Filter Architecture continuously modifies control inputs through projection to a set (dark blue) as a conservative approximation to the safe set (light blue) to ensure safety while minimizing deviation from the primary controller’s intended actions.

**Simplex Architecture** The Simplex Architecture (Seto *et al.*, 1998) employs a switching mechanism between a primary learning controller and a verified backup controller. Let $\pi : \mathcal{S} \to \mathcal{A}$ be the learned policy and $\pi_{\text{backup}} : \mathcal{S} \to \mathcal{A}$ be the backup policy. The Simplex control law is defined as:

$$a(s) = \begin{cases} \pi(s) & \text{if } s \in \mathcal{S}_{\text{safe}} \\ \pi_{\text{backup}}(s) & \text{otherwise} \end{cases}$$

where $\mathcal{S}_{\text{safe}} = \{s \in \mathcal{S} | h(s) \geq 0\}$ is the safe set defined by a Simplex monitor $h : \mathcal{S} \to \mathbb{R}$, which can be either learned or estimated. This architecture enables safe exploration by providing a fallback mechanism, separating performance and safety concerns.

Chen *et al.* (2022) design a physics-based Simplex mechanism preventing BESS SoC depletion or overload by replacing unsafe actions near safety bounds. Sun *et al.* (2024) implement action selection using voltage sensitivity-based rules to minimally modify control actions. Xia *et al.* (2022) deploy Simplex for MARL in networked microgrids using dual neural networks: Safety Evaluation Network predicting safety costs (monitor) and Action Guidance Network generating safe actions (backup controller). Their approach integrates safety directly into multi-agent SAC learning, enabling safe exploration during training and deployment.

**Safety Filter (SF)** Safety filter (Hewing *et al.*, 2020; Brunke *et al.*, 2022; Hsu *et al.*, 2023), on the other hand, *continuously* modifies control inputs to ensure safety. It is often implemented using CBFs. This approach integrates safety constraints directly into the learning process, aiming for minimally invasive interventions.

AdapSafe, proposed by Wan *et al.* (2023), exemplifies the safety filter approach in Safe RL for power system frequency control. It uses Zeroing CBFs to ensure safety: $\sup_{u_r}[L_{f_s}B(s_t) + L_{f_a}B(s_t)(a^{\text{rl}} + u_r) + \alpha(B(s_t))] \geq 0$. Here, $s_t$ is the system state, $a^{\text{rl}}$ is the RL action, $u_r$ is the safety compensation, $B(s_t)$ is the CBF function, and $L_{f_s}$ and $L_{f_a}$ are Lie derivatives corresponding to the control-affine dynamics $\dot{s} = f_s(s) + f_a(s)a$. The class-$\mathcal{K}$ function $\alpha$ is adaptively tuned:

$$\alpha = e_2 \exp(-\tan(e_3 \cdot \text{clip}(\Delta f - \Delta f_{\text{bound}}, -\frac{\pi}{2}, \frac{\pi}{2}))),$$

where $\Delta f$ denotes frequency deviation, $\Delta f_{\text{bound}}$ is the safety bound, and $e_2$ and $e_3$ are tuning parameters. The safety filter is integrated into learning via reward shaping: $r_t(s_t, a_t) - e_1\|u_r\|$ , where $r_t$ is the original reward and $e_1$ is a penalty factor.

Wang *et al.* (2023b) use Physical-Informed Safety Layer which corrects unsafe actions by solving: $\arg\min_{a_t^{\text{safe}}} \frac{1}{2}\|a_t^{\text{safe}} - a_t^{\text{rl}}\|^2$ subject to $\hat{h}^{\text{safe}}(a_t^{\text{safe}}) \geq 0.5$, where $a_t^{\text{rl}}$ is the original RL action and $a_t^{\text{safe}}$ is the

**Table 5.1:** Comparison of RTA Applications in Power Systems

| Context | RTA Architecture | Learning Integration | Implementation |
|---|---|---|---|
| (Wan *et al.*, 2023): LFC with renewables; Freq nadir, RoCoF | SF with CBF; QP-based optimization; Unlatched recovery; Self-tuning CBF parameters | Exploration and deployment; Action post-processing; Meta-learning adaptation | GP-based model adaptation; Innocuity via forward invariance |
| (Shi *et al.*, 2023): Active voltage control; Voltage bounds; MARL | SF with QP; First-order optimization; Data-driven voltage prediction | Training and deployment; Action correction penalty | 322-bus scalability; Centralized safety layer; Demonstrated nuisance-freedom |
| (Zhao *et al.*, 2023): Transient stability; Freq/voltage bounds | SF with neural CBFs; Gradient-based OPT; Unlatched filtering | DDPG primary control; Barrier pretraining phase; Online adapt. | Demonstrated nuisance-freedom; Centralized training |
| (Xia *et al.*, 2022): Microgrid freq control; Freq. and operational safety | Simplex; Prediction-correction scheme; Unlatched monitoring | Simultaneous training of safety and RL; SAC primary control; Integrated prediction-guidance | Fully decentralized; No communication needed; Safety violation tracking |
| (Sun *et al.*, 2024): VVC in unbalanced networks; V bounds | Simplex; Sensitivity-based optimization; Human-guided intervention | Training with hybrid replay; DRL primary; Human guidance in actor loss | Local sensitivity control; Voltage profile metrics; Minimal modifications |
| (Wang *et al.*, 2023b): Multi-energy microgrids; Power/gas constraints | SF-like with binary classifier; Optimization-based correction | Training and deployment safety; PPO primary; Online rule updates; Safe exploration | Multi-agent scaling; Online adaptation |
| (Zhang *et al.*, 2023c): Voltage control; V bounds | SF with DNN projection; Finite iteration algorithm | DRL primary; Integrated projection DNN; Safe space exploration | 33-bus system; Zero violations achieved |
| (Chen *et al.*, 2022): Active voltage control; V/Battery SoC constraints | Simplex; Physics-based back up actions; Local shielding | Training and deployment integration; DRL primary; Shield-guided critic training | Centralized training; Distributed execution |

corrected safe action. A supervised learning model $\hat{h}^{\text{safe}}$ is trained to classify the safety of current operating points using logistic regression, which provides a reasonably accurate approximation of the true safe operating region. The security assessment rule is continuously updated during RL training. Note that the accuracy depends on the quality and coverage of the training data and may not capture all possible constraint violations, especially for rare or extreme scenarios.

Zhao *et al.* (2023) introduce an adaptive online update mechanism to handle model uncertainties in power systems. This mechanism minimizes the following objective at each time step, effectively projecting the policy's action onto the safe set defined by the barrier function:

$$\max(0, \nabla B(s)^\top f(s, \pi_\theta(s) + u_r)) + e\|u_r\|_2^2$$

where $B(s)$ is the learned barrier function, $e$ is a weight adjustment, $\pi_\theta(s)$ is the control policy, and $u_r$ is a refinement to the control action. By adapting the control action in real-time, this approach provides a practical way to maintain safety guarantees when faced with model inaccuracies or unseen scenarios.

## 5.2 Integration of Safety into Learning

Integrating safety into learning allows policies to internalize constraints during training rather than relying on external restrictions.

Chen *et al.* (2022) store both pre- and post-filtered actions in experience replay, enabling critics to learn from safety interventions while penalty terms encourage actors to generate inherently safer actions. This coupling between safety filtering and policy learning internalizes constraints during training.

Zhang *et al.* (2023c) integrate safety constraints into DRL through a DNN-assisted projection mechanism active in both forward and backward passes. During the forward pass, the actor's action $a = \pi_\theta(s)$ is projected through a trained DNN projector $\mathcal{P}_{\text{safe}}$ to produce feasible actions $\hat{a} = \mathcal{P}_{\text{safe}}(a)$ that satisfy power system voltage constraints. In backward pass, projection gradients flow through actor updates via $\nabla_\theta J(\theta) = \mathbb{E}[-\nabla_{\hat{a}} Q(s, \hat{a}) \nabla_\theta \mathcal{P}_{\text{safe}}(\pi_\theta(s))]$, where $Q(s, \hat{a})$ is the critic's value estimate for state-action pairs and $J(\theta)$ is the policy performance objective. This architecture allows the actor to learn policies that inherently generate safe actions rather than relying on external post-hoc filtering.

Shi *et al.* (2023) combine MADDPG with trust region concepts through a data-driven safety layer for action correction. The safety layer uses first-order voltage prediction $v^*(o, a + \delta a) \approx v(o, a; \theta_v) + \nabla_a v(o, a; \theta_v) \cdot \delta a$, where $v(o, a; \theta_v)$ is the predicted voltage from neural network with parameters $\theta_v$, $o$ is the current observations, $a$ is the original action, $\Delta a$ is the action correction. Safe actions are computed via quadratic programming: $\min_{\delta a} \frac{1}{2}\|\delta a\|^2$ s.t. $0.95 \leq v(o, a; \theta_v) + \nabla_a v(o, a; \theta_v) \cdot \delta a \leq 1.05$ which finds the smallest possible action correction that ensures voltages stay within safe limits. Policy learning incorporates safe actions through trust region updates, i.e., (4.3), which ensures the policy stays close to safe actions. An action correction sub-network (ACS) learns to mimic safety layer corrections by minimizing the loss (for agent $k$): $\text{dist}(\hat{a}_k, a_k) + e\|a_k - a_k^{\text{safe}}\|^2$, where $\text{dist}(\hat{a}_k, a_k)$ is the distance function based on Q-values between the original action $\hat{a}_k$ and the corrected action $a_k$, and $a_k^{\text{safe}}$ is the safe action from safety layer. This loss guides the sub-network to learn corrections that both improve Q-values and stay close to safety layer outputs. During decentralized execution, the ACS is used in conjunction with the main policy network similar to a safety layer to produce a correction. This allows the agent to approximate the behavior of the centralized safety layer using only local information.

## 5.3 Practical Considerations and Future Directions

RTA framework has three key properties (Hobbs *et al.*, 2023). First, innocuity ensures that safety interventions preserve all system constraints. Second, viability requires that from any state, there exists a sequence of actions that maintains safety indefinitely. This implies the RTA must never allow the system to enter states where safety recovery becomes impossible. Finally, nuisance-freedom demands minimal intervention.

Nuisance-freedom has been primarily addressed through optimization approaches. For instance, Wan *et al.* (2023) employ QP formulation with self-tuning parameters, while Zhang *et al.* (2023c) explicitly minimize the $\ell_2$-norm between desired and safe actions. Innocuity and viability have received partial attention, as most works consider only instantaneous safety constraints without addressing potential future state violation. Nevertheless, works that use CBF theory to prove forward invariance, such as (Wan *et al.*, 2023), addresses both innocuity and viability. This is because the CBF condition $\sup_{a \in \mathcal{A}}[\dot{B}(s) + \alpha(B(s))] \geq 0$ ensures the existence of an instantaneous safe action, which implies the existence of infinite-horizon safe trajectories required for viability.

While some progress has been made in handling stochastic dynamics through GP regression and uncertainty handling, developing probabilistic safety guarantees remains an open challenge. The integration of formal methods with RL has seen initial steps through CBF-based approaches, but significant work remains in developing frameworks that can provide rigorous safety guarantees while accommodating RL's adaptive nature. Additionally, enhancing interpretability and explainability of RL policies through techniques such as symbolic regression remains largely unexplored in current RTA implementations. These challenges suggest promising directions for safety system architecture.

# 6

# Power System Applications

This chapter examines how Safe RL addresses critical power system problems across multiple timescales (Fig. 6.1), from sub-seconds to day-ahead planning, highlighting how operational challenges inform appropriate SRL method selection.

At fast timescales (sub-seconds to minutes), frequency regulation (Sec. 6.1) and volt-var control (Sec. 6.2) demand real-time computation and rapid response. FR maintains the critical generation-consumption balance underpinning system stability, with challenges in coordinating diverse resources under reduced system inertia while enforcing strict frequency bounds. VVC ensures power quality and voltage stability across distribution networks, growing more complex with high DER penetration and requiring coordination of numerous devices while respecting equipment limitations. Both applications face challenges from communication delays and physical safety guarantees, where SRL methods prioritize computational efficiency through safety-certified policies and action masking for both discrete and continuous control.

At intermediate timescales (minutes to hours), optimal power flow (Sec. 6.3) and critical load restoration (Sec. 6.5) balance computational complexity with system safety. OPF determines the most economic

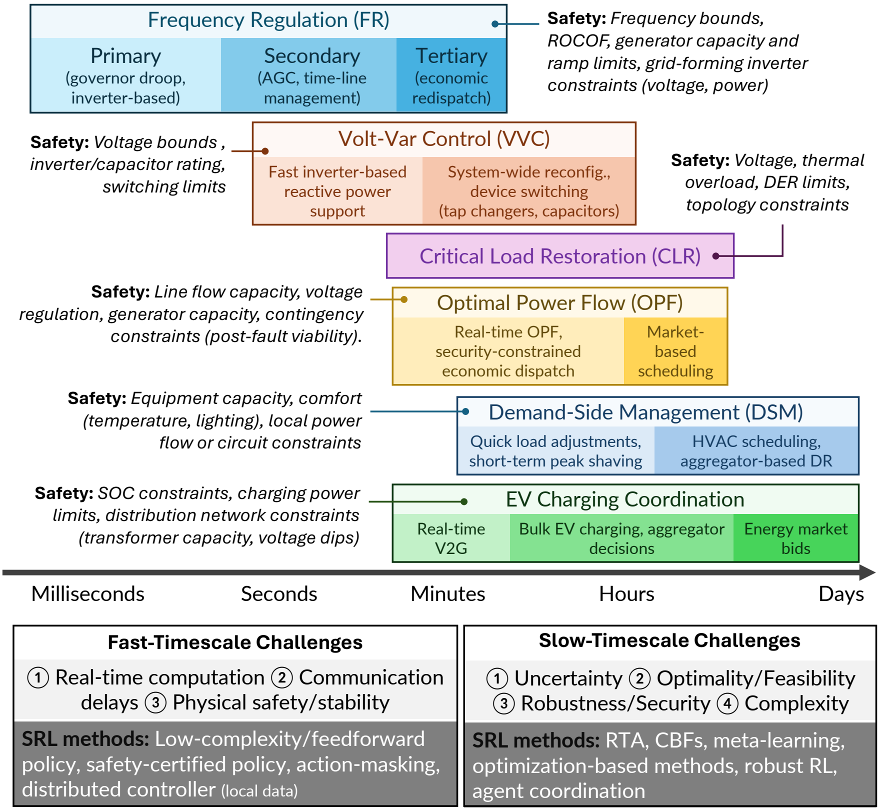

Frequency Regulation (FR)
Primary
(governor droop, inverter-based)
Secondary
(AGC, time-line management)
Tertiary
(economic redispatch)
Safety: Frequency bounds, ROCOF, generator capacity and ramp limits, grid-forming inverter constraints (voltage, power)
Safety: Voltage bounds , inverter/capacitor rating, switching limits
Volt-Var Control (VVC)
Fast inverter-based reactive power support
System-wide reconfig., device switching (tap changers, capacitors)
Safety: Voltage, thermal overload, DER limits, topology constraints
Critical Load Restoration (CLR)
Safety: Line flow capacity, voltage regulation, generator capacity, contingency constraints (post-fault viability).
Optimal Power Flow (OPF)
Real-time OPF, security-constrained economic dispatch
Market-based scheduling
Safety: Equipment capacity, comfort (temperature, lighting), local power flow or circuit constraints
Demand-Side Management (DSM)
Quick load adjustments, short-term peak shaving
HVAC scheduling, aggregator-based DR
Safety: SOC constraints, charging power limits, distribution network constraints (transformer capacity, voltage dips)
EV Charging Coordination
Real-time V2G
Bulk EV charging, aggregator decisions
Energy market bids
Milliseconds
Seconds
Minutes
Hours
Days
Fast-Timescale Challenges
① Real-time computation ② Communication delays ③ Physical safety/stability
SRL methods: Low-complexity/feedforward policy, safety-certified policy, action-masking, distributed controller (local data)
Slow-Timescale Challenges
① Uncertainty ② Optimality/Feasibility ③ Robustness/Security ④ Complexity
SRL methods: RTA, CBFs, meta-learning, optimization-based methods, robust RL, agent coordination

**Figure 6.1: Overview of Power System Applications by Timescale and Safe RL Challenges.** The top portion arranges key applications along a time axis (horizontal bar), from fast frequency regulation (sub-seconds to minutes) through volt–var control and critical load restoration (minutes to hours), up to optimal power flow, demand-side management, and EV charging coordination (hours to day-ahead). Each application notes its key safety constraints. The lower labels highlight major challenges across different timescales and give examples of Safe RL methods (RTA, CBFs, meta-learning, action masking, distributed controllers) that address these challenges at different scales.

operating point while satisfying complex network constraints and operational limits under significant generation and demand uncertainties. CLR ensures rapid recovery after outages through coordinated resource management. These applications require coordinating multiple devices under uncertainty, where SRL approaches employ control barrier functions and RTAs to maintain safety constraints while optimizing performance.

At slower timescales (hours to days), demand-side management (Sec. 6.4) and EV charging coordination (Sec. 6.6) address large-scale optimization under uncertainty. DSM leverages flexible loads for system efficiency and renewable integration, balancing grid objectives with user comfort while respecting system and device limitations. EV charging coordination manages growing vehicle-grid interactions, satisfying charging requirements while preventing network congestion. These applications contend with extensive state-action spaces and extended planning horizons, where multi-agent methods excel for geographically distributed systems.

Beyond these operational controls, power system state estimation (Sec. 6.7) provides foundational situational awareness by handling model uncertainties and measurement anomalies with bounded error guarantees. Cybersecurity (Sec. 6.8) enables adaptive detection and mitigation of sophisticated threats while maintaining stability during attacks. These applications demonstrate Safe RL's adaptability to specialized power system requirements.

This chapter builds upon the fundamental Safe RL algorithms from Chapters 2 to 5. To illustrate how theoretical frameworks can be translated into practical power system control solutions while maintaining safety guarantees, we provide detailed formulations for two key applications: FR and VVC. Throughout this chapter, we highlight both the commonalities in safe learning across applications and the unique considerations that arise in each domain. The subsequent sections examine each application in detail, supported by comprehensive comparisons of recent research advances in Tables 6.1–6.6.

## 6.1 Frequency Regulation

Frequency regulation maintains the critical balance between power generation and consumption, with deviations from nominal values potentially causing equipment damage, instability, or blackouts. Traditional control architecture operates across three timescales: primary control (seconds, via governors and grid-forming inverters), secondary control/AGC (minutes, restoring nominal frequency and managing tie-lines), and tertiary

**Table 6.1:** Comparison of SRL for Frequency Control

| Control Level | System Model | Safety Types | Key Assumptions | Test Systems |
|---|---|---|---|---|
| (Jin and Lavaei, 2020): Primary control | Swing equation dynamics | Stability certificates via robust control IQCs | Uncertain but bounded system parameters, Small angle assumption | IEEE 39-bus system, Initial condition responses tested |
| (Xia *et al.*, 2022): Multi-level control, Economic optimization | Multi-source integration, Linearized LFC model, Area-based control | Frequency bounds, ESS constraints, Safety prediction | Small signal stability, Decentralized control, Known system parameters | Multi-microgrid network, Real load/PV profiles, Various scenarios |
| (Gu *et al.*, 2022): Primary control | Linearized synchronous generator model, Swing equation dynamics | Exponential stability via Lyapunov certificates | Uncertain but bounded system parameters, Linearized operation | IEEE NE39, Single communication topology variants |
| (Cui *et al.*, 2023): Primary control, Local frequency regulation | SG and IBR integration, PLL-based inverters, Swing equation dynamics | Frequency limits, Power constraints, Local Lyapunov stability | Lossless system, Known inertia and damping, Angle difference bounds | IEEE NE39, Step load changes, Kron reduction, Random conditions |
| (Wan *et al.*, 2023): Integrated primary and secondary control | Swing equations, Governor-droop model | Frequency nadir, ROCOF limits, Power bounds, CBF | Uncertain but bounded parameters, Linear load damping model | GB 2030 system, Multiple contingencies, Parameter variations |
| (Kwon *et al.*, 2023): Primary control, Fast frequency regulation | SG-GFM hybrid, P-$\omega$ & Q-V droop control, Network-coupled model | Risk-constrained LQR, Frequency stability, Cost variance bounds | Linearized model, Neighbor communication, Parameter certainty | Modified IEEE 68-bus, GFM integration, Multiple scenarios |
| (Zhao *et al.*, 2023): Primary transient stability control | Swing equations for SGs, Droop-controlled inverter dynamics | Frequency/Voltage limits, Transient stability, CBFs | Interface voltage constant in transients | IEEE 13-, 39-, 118-bus systems, Transient stability scenarios |
| (Shuai *et al.*, 2024): Primary control, Fast regulation | GFM-based system, Swing equation dynamics, VSG control, AC power flow | Power limits, Frequency Lyapunov stability, Disturbance handling | Lipschitz dynamics, Neglected reactive power, Parameter uncertainty | Single GFM setup, Progressive disturbances |
| (Liu *et al.*, 2024b): Secondary control, Tie-line management | Traditional SG focus, Multi-area model, Classic dynamics | Frequency stability, Lyapunov guarantees, Attack robustness | Small-signal model, Available ACE measurements, Known system parameters | IEEE 39-bus, FGSM cyber attacks |
| (Yuan *et al.*, 2024): Transient frequency control | Swing equations, Aggregate bus model | Frequency bounds, Lyapunov stability | Known system parameters, Disturbance vanishes in finite time | IEEE 39-bus, Measurement noise tests, Partial info scenarios |

control (15-30 minutes, optimizing generation and restoring reserves). Conventional approaches using proportional-integral (PI) controllers and droop control face challenges from renewable energy uncertainty and reduced system inertia.

RL for FR demonstrates several technical approaches:

Control architecture implementations span primary to secondary layers. Fast-timescale approaches (Cui *et al.*, 2023; Shuai *et al.*, 2024) focus on local frequency regulation through grid-forming (GFM) inverters, operating at millisecond timescales. Secondary control frameworks address Area Control Error (ACE) and tie-line management (Liu *et al.*, 2024b). Some approaches bridge timescales—integrating inertial response with primary/secondary objectives (Wan *et al.*, 2023) or combining frequency regulation with economic dispatch (Xia *et al.*, 2022).

System modeling reflects modern grid requirements, incorporating Phase-Locked Loop (PLL) dynamics (Cui *et al.*, 2023), voltage-source converter control (Kwon *et al.*, 2023), and the swing equation at varying complexity levels. Safety frameworks employ CBFs for explicit constraints on frequency nadir and ROCOF (Wan *et al.*, 2023) and risk-constrained LQR for stability under uncertainty (Kwon *et al.*, 2023).

Parameter uncertainty treatments include: known parameters with local measurements (Cui *et al.*, 2023), explicitly bounded uncertainties (Wan *et al.*, 2023), and Gaussian Process modeling for unknown dynamics (Shuai *et al.*, 2024). The treatment of uncertainty becomes particularly relevant for renewable integration scenarios where system parameters may vary significantly.

Test implementations span single-GFM setups to modified IEEE cases, with validation focusing on frequency response under various disturbances. Load changes, generator outages, and cyber attacks serve as common test scenarios. Some studies incorporate explicit communication constraints (Kwon *et al.*, 2023; Liu *et al.*, 2024b).

### 6.1.1 Safe RL Formulation for Frequency Control

The frequency control problem can be formulated as a CMDP:

**State Space** The state space $\mathcal{S}$ captures power system dynamics, e.g., $\{\Delta f, \frac{d\Delta f}{dt}, p_{\text{gen}}, p_{\text{tie}}, p_{\text{load}}^{\text{hist}}, p_{\text{PV}}^{\text{hist}}\} \subseteq s \in \mathcal{S}$, where $\Delta f$ is frequency deviation from nominal value (Hz), $\frac{d\Delta f}{dt}$ is RoCoF (Hz/s), $p_{\text{gen}}$ is generator outputs, $p_{\text{tie}}$ represents tie-line flows between interconnected areas, and $p_{\text{load}}^{\text{hist}}, p_{\text{PV}}^{\text{hist}}$ capture historical load and renewable generation. The state space design reflects a fundamental trade-off between observability and complexity. For example, including historical measurements helps handle system delays and uncertainties, while the RoCoF term enables better dynamic response prediction. The specific choice of state variables may vary based on the application context—interconnected transmission systems might emphasize tie-line flows, while isolated microgrids might focus more on local power balance.

**Action Space** The action space $\mathcal{A}$ comprises control adjustments, e.g., $\{\Delta p_{\text{gen}}, \Delta p_{\text{v}}, \Delta p_{\text{ESS}}\} \subseteq a \in \mathcal{A}$, where $\Delta p_{\text{gen}}, \Delta p_{\text{v}}, \Delta p_{\text{ESS}}$ represent adjustments to conventional generators, renewable sources, and energy storage respectively. The action space design should account for the characteristics of different power sources. For instance, conventional generators have slower response times but larger capacity, while inverter-based resources offer faster response but may have limited power reserves.

**Reward Function** The reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ balances multiple control objectives. For example, $r(s, a) = -(e_1\|\Delta f\|^2 + e_2\|a\|^2 + e_3\|p_{\text{tie}} - p_{\text{tie}}^{\text{ref}}\|^2 + e_4 c_{\text{gen}})$, where $e_1, ..., e_4$ are weighting coefficients, $\|\Delta f\|^2$ penalizes frequency deviations, $\|a\|^2$ represents control effort, $\|p_{\text{tie}} - p_{\text{tie}}^{\text{ref}}\|^2$ maintains scheduled tie-line flows, and $c_{\text{gen}}$ represents generation costs as a function of $p_{\text{gen}}$. For multi-agent systems, the reward can include cooperation terms, e.g., $r_k = r_k^{\text{local}} + e_{\text{coord}} \sum_{k' \in \mathcal{N}(k)} r_{kk'}^{\text{coupling}}$, where $\mathcal{N}_k$ represents the neighboring areas and $e_{\text{coord}}$ balances local and cooperative objectives.

**Safety Constraints** Safety in frequency control involves both operational limits and stability guarantees. The operational constraints include frequency limits ($f_{\min} \leq f \leq f_{\max}$), RoCoF bounds ($|\frac{d\Delta f}{dt}| \leq \Delta \dot{f}_{\text{bound}}$), generation limits ($p_{\text{gen,min}} \leq p_{\text{gen}} \leq p_{\text{gen,max}}$), ramp rate constraints ($|\frac{dp_{\text{gen}}}{dt}| \leq r_{\max}$). Stability guarantees can be implemented through multiple complementary approaches, including CBFs (Wan *et al.*, 2023) and Lyapunov stability (Cui *et al.*, 2023).

The choice of safety implementation depends on system requirements and computational resources. These constraints can be encoded as expected cumulative constraint for long-term averages, which is suitable for managing thermal limits or battery lifecycles, or almost surely instantaneous constraint for strict bounds, essential for frequency limits. CBFs provide continuous safety guarantees but may be computationally intensive, while learned safety models offer computational efficiency but may provide weaker guarantees.

**Performance Evaluation Metrics** The effectiveness of RL controllers is evaluated using several key metrics.

Frequency-related metrics include mean absolute frequency deviation: $\sum_t |\Delta f(t)|$, peak frequency deviation: $\max_t |\Delta f(t)|$, and maximum RoCoF: $\max_t |\frac{d\Delta f}{dt}(t)|$, and frequency nadir: $\min_t f(t)$, i.e., lowest point that the system frequency reaches following a disturbance before it starts to recover. The frequency nadir captures transient stability concern: if the frequency falls too low during this transient period, protective relays may trigger, leading to cascading outages and potential blackouts. The system needs to maintain synchronism through this critical period.

The control effort can be evaluated through average control effort: $\sum_t \|a(t)\|^2$, generation costs: $\sum_t c_{\text{gen}}(p_{\text{gen}}(t))$. For interconnected systems, the performance can be evaluated using tie-line power deviation: $\sum_t \|p_{\text{tie}}(t) - p_{\text{tie}}^{\text{ref}}\|^2$, and local area control error: $\sum_t \|ACE_k(t)\|^2$.

For safety constraints, $\sum_t \mathbb{I}(c(s_t, a_t) > \xi_t)$ counts constraint violations. For stability, the region of attraction is typically examined (Gu *et al.*, 2022; Cui *et al.*, 2023; Shuai *et al.*, 2024).

### 6.1.2 Safe RL Techniques for Frequency Regulation

**Lyapunov-Based Stability Methods** These methods incorporate model-based Lyapunov stability analysis into learning, providing rigorous stability guarantees through mathematical certificates. Jin and Lavaei (2020) establish this approach by combining TRPO with stability certificates based on IQCs, implementing stability penalties for gradient regulation and hard thresholding of network weights. This provides input-output stability guarantees while maintaining learning performance through careful regulation of the policy update process. Cui *et al.* (2023) introduce RNN-based controllers with Lyapunov certification, enforcing monotonicity constraints while providing local exponential stability guarantees. Shuai *et al.* (2024) combine MBRL with Lyapunov theory and approximate dynamic programming, using a dual-purpose Lyapunov function for determining attraction regions and optimization, with Gaussian Process modeling for uncertainty quantification. Liu *et al.* (2024b) integrate DDPG with monotonic neural networks, enforcing deviation-command monotonicity through architecture rather than external constraints. Yuan *et al.* (2024) advance this category with a dynamic budget mechanism within the Lyapunov framework, combining RNN-based architecture with distributed safety enforcement that allows temporary local violations while maintaining global stability.

**Barrier Function-Based Methods** These approaches utilize CBFs to ensure safety through forward invariance properties of safe sets. They typically use safety filter as the architecture and integrate with model-free RL. Zhao *et al.* (2023) introduce a neural barrier function approach combining DDPG with a barrier-certification system, implementing a two-stage process where control actions are filtered through neural barrier functions, providing bounded generalization error guarantees and adaptive refinement for online operation. Wan *et al.* (2023) present an adaptive CBF-based framework with meta-learning capabilities, implementing a hierarchical structure where a CBF compensator modifies RL policy actions, incorporating adaptive Gaussian Process regression to learn model corrections online and address model uncertainty.

**Risk-Constrained Methods** These methods enhance robustness for worst-case scenarios through statistical risk measures or adversarial training. Kwon *et al.* (2023) develop a risk-constrained approach integrated with LQR, explicitly bounding state cost variability through Lagrangian relaxation. The solution method employs a stochastic gradient descent with max-oracle (SGDmax) algorithm, utilizing zero-order policy gradient (ZOPG) for efficient gradient estimation. The approach prioritizes statistical risk measures over hard constraints, focusing on robustness against uncertainties and disturbances. Liu *et al.* (2024b) incorporate Fast Gradient Sign Method (FGSM) attacks into training, creating an adversarial regime that enhances controller robustness against such attacks.

**Projection-Based Methods** These methods ensure safety through projection operations onto sets of stabilizing controllers, often utilizing convex optimization techniques. For example, Gu *et al.* (2022) develop a projected policy gradient method that enforces stability conditions through projection onto a convex set of stabilizing parameters, providing exponential stability guarantees through Lyapunov certificates, with safety constraints enforced through a projection step formulated as a semidefinite program.

**Multi-Agent Methods** Multi-agent approaches address scale by balancing local control with system stability. Xia *et al.* (2022) use a two-phase framework with CTDE. The method employs dual networks: a safety evaluation network predicting constraint violations and an action guidance network providing corrective actions. The method implements both hard constraints for physical limits and soft penalties for economic objectives, with a mechanism for selecting safe actions when violations are predicted. Each microgrid agent operates independently while meeting system stability requirements. Kwon *et al.* (2024) extend the SGDmax algorithm with zero-order policy gradient from (Kwon *et al.*, 2023) to create a distributed control framework that leverages coherency information. The key insight is that replacing traditional generators with grid-forming inverters creates predictable changes in how generators naturally group together during oscillations. The controller design exploits this by mapping the new coherent groupings after inverter integration, implementing control actions only between groups, and targeting inter-area oscillations specifically. This coherency-aware approach demonstrates better performance than traditional methods that don't account for inverter-induced changes in system dynamics.

While each approach has distinct characteristics, there is a clear trend toward hybrid methods combining traditional control theory with modern learning techniques.

### Training Approaches

Existing approaches predominantly employ offline training (Cui *et al.*, 2023; Wan *et al.*, 2023; Kwon *et al.*, 2023), generating and learning from simulated trajectories before testing. In contrast, Gu *et al.* (2022) implement online training with stability-preserving projection steps during learning. Several methods utilize multi-phase strategies: Xia *et al.* (2022) combine centralized offline training with decentralized execution, while Wan *et al.* (2023) employ meta-learning with separate meta-training and adaptation phases to validate safety properties before testing.

Safety mechanisms during training vary across implementations. Liu *et al.* (2024b) embed safety directly into the architecture through monotonic neural networks, Jin and Lavaei (2020) combine stability penalties with weight thresholding, and Xia *et al.* (2022) use a 200-episode pre-training period for safety models. Continuous safety enforcement approaches include Shuai *et al.* (2024)'s strategic exploration within Lyapunov-verified safe regions and Gu *et al.* (2022)'s parameter projection onto stabilizing controller sets after each gradient step.

Training data generation follows specific constraints: Cui *et al.* (2023) use simulated trajectories with tightly controlled parameters (initial angles between [-0.05, 0.05] rad, frequencies between [-0.1, 0.1] Hz), Kwon *et al.* (2023) generate data under varying conditions (±0.5 to ±1 pu perturbations), Xia *et al.* (2022) incorporate real-world measurements (10 hours of smart meter data), and Yuan *et al.* (2024) operate within specific bounds (initial frequencies in [59.9, 60.1] Hz, power variations of ±10%).

## 6.1.3 Practical Considerations

**Communication Requirements** Communication architectures range from minimal local measurements (e.g., frequency and voltage phase (Shuai *et al.*, 2024; Cui *et al.*, 2023)) to comprehensive state information including tie-line flows (Liu *et al.*, 2024b) and neighboring bus states (Yuan *et al.*, 2024). Inter-area communication spans from fully decentralized approaches (Xia *et al.*, 2022; Cui *et al.*, 2023) to distributed architectures requiring structured neighbor communication

(Kwon *et al.*, 2023; Jin and Lavaei, 2020). Update frequencies vary from high-frequency 100Hz updates (Kwon *et al.*, 2023) to longer 30-second intervals (Xia *et al.*, 2022).

**Computational Aspects** Computational requirements range from static control laws (Cui *et al.*, 2023) to more demanding approaches with quadratic programming and Gaussian Process regression (Wan *et al.*, 2023). Runtime performance spans from 0.01s execution (Xia *et al.*, 2022) to extended training periods, typically on standard CPU platforms (Xia *et al.*, 2022; Shuai *et al.*, 2024). Scalability concerns are addressed through decentralized implementations (Cui *et al.*, 2023) or distributed cooperation frameworks (Yuan *et al.*, 2024).

**Control Architecture and System Modeling** Frequency control operates across hierarchical layers at different timescales. Wan *et al.* (2023) explicitly group inertial response and primary control (immediate stability mechanisms) into Phase I, with slower secondary/tertiary controls in Phase II, allowing targeted safety constraints for each phase.

Inverter modeling follows two approaches: grid-forming inverters as voltage sources with specified magnitude and angle, such as P-$\omega$ and Q-V droop control (Kwon *et al.*, 2023), essential for low-inertia grids, and grid-following inverters as current sources synchronizing with existing frequency through phase-locked loops (Cui *et al.*, 2023), suitable for integration into strong grids with substantial synchronous generation. This modeling distinction is particularly crucial for low-inertia grids where the choice between grid-forming and grid-following capabilities significantly influences system stability and control performance. Grid-forming inverters become increasingly essential in systems with reduced conventional generation, while grid-following inverters remain suitable for integration into strong grids with substantial synchronous generation.

**Integration Considerations** Integration focuses on compatibility and modification requirements. Methods typically interface with existing droop control by modifying setpoints rather than replacing controllers (Cui *et al.*, 2023; Kwon *et al.*, 2023). System modifications range from measurement system installations (Xia *et al.*, 2022) to inverter interface modifications (Cui *et al.*, 2023). Backward compatibility is maintained by preserving control structures, exemplified by Kwon *et al.* (2023)'s retention of basic droop control with a higher-level coordination layer.

Certification requires hardware-in-the-loop simulations and field trials to verify performance under various disturbances and demonstrate constraint satisfaction. Compliance with grid codes (e.g., NERC BAL-003-1, ENTSO-E) requires meeting technical specifications like frequency restoration within 15 minutes and primary response activation within 20-52 seconds, alongside mathematical guarantees for constraint satisfaction.

## 6.2 Volt-Var Control

High penetration of DERs has made voltage stability and power quality management increasingly challenging in modern distribution networks. Traditional deterministic approaches prove inadequate amid growing uncertainties, motivating the development of safe RL approaches for VVC (Table 6.2).

Recent research demonstrates diverse approaches in both formulation and implementation. Model-free methods eliminate the need for accurate system models while using CMDP for safety constraints (Wang *et al.*, 2019). Model-augmented approaches enhance learning by leveraging physical insights, implementing quadratic programming-based constraint layers (Gao and Yu, 2022). Distributed architectures enable scalable control through coordinated local actions (Yan *et al.*, 2023; Shi *et al.*, 2023).

Safety mechanisms range from theoretical guarantees to practical implementations. Feng *et al.* (2023) provide global asymptotic stability through Lyapunov analysis, while Chen *et al.* (2022) implement physics-based shielding for battery systems. These frameworks prove crucial during both learning and deployment phases, with validation results demonstrating high reliability.

Implementation challenges are evident in test system variations (Table 6.2). Most methods demonstrate scalability from small IEEE test feeders to larger networks but differ significantly in communication requirements and computational efficiency. For example, Zhang *et al.* (2023b) achieve communication-free operation, while Sun *et al.* (2024) assume sufficient infrastructure for real-time coordination. Computational performance varies from milliseconds (Feng *et al.*, 2023) to minutes, highlighting the trade-off between control sophistication and real-time applicability.

### 6.2.1 Safe RL Formulation for VVC

**State Space** State space selection depends on control objectives, available measurements, system architecture, and controllable device types. The basic form involves $\{p^{\text{real}}, p^{\text{react}}, |v|, \angle v, X_{\text{status}}, X_{\text{temp}}\} \subset s \in S$, where $p^{\text{real}}$ and $p^{\text{react}}$ represent real and reactive power injections, respectively, $|v|$ and $\angle v$ represent voltage magnitudes and phase angles, respectively, $X_{\text{status}}$ represents control device states, and $X_{\text{temp}}$ represents temporal features. This suits traditional VVC devices where state history helps limit switching operations. For multi-area control, the formulation extends to area-specific information through $\{p_k^{\text{real}}, p_k^{\text{react}}, v_k\} \subset s_k \in \mathcal{S}_k$, where $p_k^{\text{real}}, p_k^{\text{react}}$ represent outlet powers of $k$-th area, enabling distributed control with limited observability (Liu and Wu, 2021). Systems with renewables can incorporate generation and uncertainty measures directly (Nguyen and Choi, 2022; Zhang *et al.*, 2023b).

**Action Space** Action space formulation varies based on controllable devices. Traditional VVC systems use discrete actions, such as tap/on-off positions for controllable devices (Wang *et al.*, 2019). Systems with inverter-based resources use continuous actions, such as reactive power ratios or set points.

**Reward Function** The reward function design for VVC incorporates several operational objectives through distinct terms. Power loss $p_{\text{loss}}(t) = \sum_l \omega_l^{\text{resist}} |I_l|^2$ computes the total real power loss across network branches, with $\omega_l^{\text{resist}}$ being the resistance and $I_l$ the current magnitude of branch $l$. Device switching costs $\sum_k |\Delta x_k|$ prevent excessive device operations, where $\Delta x_k$ represents the change in device $k$'s state between consecutive time steps. Voltage regulation $-\sum_k |v_k - v_{\text{ref}}|$, maintains voltage

**Table 6.2:** Comparison of Safe RL Approaches for VVC

| Problem Setup | System Model | Safety Aspects | Test Systems |
|---|---|---|---|
| (Wang *et al.*, 2019): Device wear reduction; Controls voltage regulators, tap changers, capacitors | Model-free implementation; Centralized architecture | Constrained soft actor-critic; Local optimality guarantees; Voltage violation prevention | IEEE 4, 34, 123-bus feeders; One year London smart meter data |
| (Liu and Wu, 2021): PV inverters and SVC control; Online decentralized framework; Fast-timescale operation | Balanced steady-state model; Communication delay; Asynchronous implementation | Safe exploration mechanism; Exploration-exploitation balance; Voltage limits | IEEE 33, 141-bus balanced; IEEE 37-bus unbalanced; Convergence validation |
| (Gao and Yu, 2022): Tap and capacitor control; Model-augmented RL; Probabilistic uncertainty handling | DistFlow equations; Probabilistic NN for uncertainty; SCADA/AMI data required | QP constraint layer; MI regularizer; Empirical validation; Iterative QP satisfaction | IEEE 4, 34, 123-bus systems; London meter data |
| (Feng *et al.*, 2023): Transient performance focus; DER reactive power control; Decentralized; Fast violation response | Branch flow model; Linear approximation; Fast reactive power loop; Balanced phase assumption | Lyapunov stability analysis; Global asymptotic guarantees; Monotonic policy | IEEE 13, 123-bus systems; Single/three-phase versions; Fast computation (0.37ms) |
| (Wang *et al.*, 2023b): Multi-energy microgrid management; Cost minimization; DG/storage/gas control | AC power flow with gas model; Unknown parameters; Price uncertainties; Power-gas coupling | Physical-informed safety layer; Voltage/thermal/pressure limits | 6-bus power, 7-node gas; 33-bus power, 20-node gas |
| (Zhang *et al.*, 2023b): VCC and loss minimization; Smart converter control; Projection-based safety | Branch power flow model; Steady-state operation; RES uncertainty handling | Projection layer for safety; Voltage constraints; Zero violations during training | Modified IEEE 33-bus; 6 PV converters; 8 variable loads; 30-min resolution data |
| (Yan *et al.*, 2023): Multi-zone VVC; PV inverter reactive power control; Decentralized framework | AC power flow; Limited zone information; Normal distribution uncertainty | Voltage limits; Primal-dual optimization; Empirical safety validation | 141-bus system; 9 zones; Efficient computation (81ms) |
| (Sun *et al.*, 2024): Three-phase unbalance compensation; PV inverter coordination | Three-phase unbalanced model; Simplified voltage equations | Voltage constraints; Hybrid experience replay; Sensitivity-based guidance | Modified IEEE 123-bus; 42 single-phase PV inverters |
| (Nguyen and Choi, 2022): Three-stage VVC; Peak reduction; Voltage regulation; Hierarchical framework | Full and linear models; Perfect communication; Radial topology; Fast local measurements | Three-layer safety structure; Day-ahead stability; Real-time adjustments | IEEE 33, 123-bus systems; Multi-timescale validation |
| (Chen *et al.*, 2022): Active voltage control; BESS safety focus; Power congestion management; Decentralized execution | Steady-state balanced model; Control at PV buses; Gaussian uncertainty | Physics-based shield mechanism; SoC protection; Voltage limits | IEEE 33, 141-bus systems; Three-year Portuguese data; Three-minute control period |
| (Shi *et al.*, 2023): Active voltage control; Loss minimization; PV inverter coordination; CTDE | Nonlinear AC power flow; Steady-state model; Measurement uncertainty | Data-driven safety layer; Action correction mechanism; First-order approximation guarantees | 33-bus (6 PVs), 141-bus (22 PVs), 322-bus (38 PVs); Three years real-world data |
| (Guo *et al.*, 2023): High PV penetration; Multi-agent decentralized control | Full AC power flow; Second order cone model; Branch flow constraints | Safety projection; State synchronization block; Delay-robust operation | IEEE 33-bus system; 15-min resolution |
| (Jeon *et al.*, 2023): Multi-objective VVC; EV charging support; MESS coordination | DistFlow equations; ZIP load model; Coupled transportation network | SOC/Voltage safety modules; Iterative violation correction; Plug-and-play | IEEE 33- and 57-bus with 15- and 42-node transport |

profiles, where $v_k$ is the voltage magnitude at bus $k$, and $v_{\text{ref}}$ is the reference voltage (typically 1.0 p.u.). Economic formulations incorporate time-varying electricity prices through $c(t)$, modifying the loss term to

$-c(t)p_{\text{loss}}(t)$. Systems with renewable integration often include reactive power utilization costs $-\sum_k |p_k^{\text{react}}|$.

**Safety Constraints** Safety constraints in VVC primarily focus on maintaining voltage stability. The fundamental constraint $v_{\min} \leq |v_k| \leq v_{\max}$ ensures voltage magnitudes remain within acceptable bounds, typically $v_{\min} = 0.95$ p.u. and $v_{\max} = 1.05$ p.u. Equipment constraints include apparent power limits and ramp rate limitations $|\Delta p_k^{\text{react}}| \leq \Delta p_{k,max}^{\text{react}}$ to prevent rapid changes in reactive power output.

**Performance Evaluation Metrics** Technical performance is measured by power loss reduction $\Delta p_{\text{loss}}$ relative to baseline operation, voltage profile improvement representing the standard deviation of voltage magnitudes, and control action count tracking device operations. Safety compliance is evaluated through constraint violation frequency and maximum deviation magnitude $\max_k |v_k - v_{\text{ref}}|$. Learning efficiency metrics assess convergence rate, sample efficiency, and average decision time.

**Three-Phase Considerations** Three-phase VVC extends single-phase control by addressing phase imbalances and interactions. The methods generally follow two strategies: integrated three-phase modeling and phase-decoupled control. Integrated modeling (e.g., (Sun *et al.*, 2024)) incorporates phases in state-action spaces through phase-specific power injections, loads, and voltages, with phase-wise reactive power control. Three-phase voltage sensitivity matrices capture cross-phase effects, while Voltage Unbalance Factor (VUF) constraints manage imbalances. Phase coupling through impedance matrices enables coordinated control.

Phase-decoupled control (e.g., (Feng *et al.*, 2023)) treats phases separately using diagonal control structures that preserve monotonicity per phase. States can be arranged by bus or phase, with independent controllers operating under system-wide stability conditions. This reduces computational complexity while retaining control effectiveness.

Performance evaluation focuses on voltage profile improvement across phases, reduction in phase imbalance, and control efficiency.

### 6.2.2 Safe RL Techniques for VVC

**SAC-Based Methods** These methods adapt SAC to incorporate safety constraints, leveraging maximum entropy principles and off-policy training for voltage control challenges. The core elements include dual critics for value estimation, entropy tuning, and various mechanisms for constraint satisfaction (see Sec. 2.1.3 for technical details). Wang *et al.* (2019) formulate VVC as a CMDP through CSAC, using Lagrangian multipliers for voltage constraints and dual gradient descent for policy updates. Sun *et al.* (2024) implement a voltage sensitivity-based intervention module with a hybrid experience replay buffer storing both safe and unsafe transitions. Nguyen and Choi (2022) develop a safety module using iterative voltage control equation $a_{k,t}(i+1) = a_{k,t}(i) + \rho(a_{j,t}^{agent} - a_{k,t}(i))$, essentially a safe gradient-based update between the RL agent's desired reactive power output $(a_{j,t}^{agent})$ and the current reactive power injection $(a_{j,t}(i))$ at node $j$, time $t$, and iteration $i$. The adaptive parameter $\rho$ modulates the update speed based on voltage proximity to limits—it increases when voltages are well within bounds (allowing faster convergence to agent's action) and decreases as voltages approach limits (enforcing more conservative updates). Jeon *et al.* (2023) implement dual safety modules for SOC and voltage constraints, using adaptive parameters that increase correction strength when approaching constraint boundaries.

**RTA-Based Methods** RTA methods provide real-time safety guarantees through optimization-based action modification or monitoring-based intervention, working independently or with learning algorithms such as SAC (Nguyen and Choi, 2022; Jeon *et al.*, 2023). Gao and Yu (2022) implement a QP-based safety layer minimizing deviation from the learned policy while ensuring voltage constraints. Zhang *et al.* (2023c) propose DNN Projection embedded twin-delayed deep deterministic policy gradient (DPe-TD3), integrating a finite iteration projection algorithm for hard constraints and a DNN-assisted projection layer for computational efficiency. Shi *et al.* (2023) develop a centralized safety layer using first-order voltage predictions for efficient action correction. Wang *et al.* (2023b) integrate security assessment with PPO, solving

$\min \|a_t^{safe} - a_t^{ppo}\|^2$ subject to probabilistic security constraints, where actions are modified to maintain system security while minimizing deviation from the PPO policy.

**Multi-Agent Methods** Multi-agent methods enable coordinated voltage control while maintaining system-wide safety constraints, using either projection-based mechanisms or explicit constraint formulations. Liu and Wu (2021) extend SAC to multi-agent settings through MAC-SAC, using constrained Markov games with individual agent constraints on voltage expectations. Zhang *et al.* (2023b) implement MADDPG with projection within each agent's actor network to ensure feasible actions. Yan *et al.* (2023) use GCNs for network-aware feature extraction and primal-dual optimization for constraint satisfaction in a decentralized framework. Chen *et al.* (2022) implement MATD3 with physics-based shields modifying actions according to power system constraints and battery characteristics. Guo *et al.* (2023) implement an analytical safety projection layer based on branch power flow models, which minimizes control adjustment subject to explicit voltage constraints and thermal limits. Kabir *et al.* (2023) propose a hierarchical two-timescale approach using MASAC for slow-timescale conventional devices and DDPG for fast-timescale smart inverter control, demonstrating how CTDE can address non-stationarity in two-timescale voltage control.

### Training Approaches

Training approaches in safe RL for VVC reflect key power system considerations. Online versus offline training choices stem from voltage stability requirements during learning. Wang *et al.* (2019) and Sun *et al.* (2024) use offline training, while Liu and Wu (2021) and Yan *et al.* (2023) implement online learning with constraint handling through dual variable updates and primal-dual optimization. For MARL, CTDE addresses the trade-off between coordinated control and communication constraints (Zhang *et al.*, 2023b; Guo *et al.*, 2023; Shi *et al.*, 2023), enabling coordinated policy learning through shared experience while allowing real-time control using local measurements.

Data requirements address power system characteristics, typically spanning 1-3 years of load profiles with 1-15 minute resolution. Chen *et al.* (2022) and Zhang *et al.* (2023c) augment real data with synthetic samples by adding Gaussian noise to PV generation and load profiles. Experience collection strategies differ—Sun *et al.* (2024) and Guo *et al.* (2023) store both safe and unsafe transitions in hybrid buffers, while Gao and Yu (2022) train only on projected safe actions. Hu *et al.* (2022b) address sample efficiency through Experience Augmentation, exploiting distribution system symmetry to generate synthetic training data while accelerating convergence toward safe voltage control policies.

### 6.2.3 Practical Considerations

**Control Architecture Evolution** Multi-layered control structures balance system-wide optimization with local responsiveness. Nguyen and Choi (2022) implement a three-stage framework coordinating day-ahead, real-time, and local timescales. Yan *et al.* (2023) examine zone-based architectures established in European networks, while Zhang *et al.* (2023b) introduce system partitioning using sensitivity matrices and spectral clustering to reduce subsystem coupling.

**Resource Integration and Coordination** VVC systems coordinate diverse grid technologies with varying characteristics. Liu and Wu (2021) examine inverter-based resources providing support using free capacity, while Jeon *et al.* (2023) analyze mobile energy storage integration considering both power system and transportation constraints. Chen *et al.* (2022) demonstrate joint active-reactive power control effectiveness due to distribution system X/R ratio characteristics. Wang *et al.* (2023b) examine integrated management of power and gas networks through gas-fired generators as coupling points, enabling coordinated dispatch while considering physical constraints of both networks.

**Model Uncertainty and Data Management** Control approaches adapt to handle model uncertainty and data requirements. Wang *et al.* (2019) note utilities' challenges in maintaining network models across millions of buses. Gao and Yu (2022) develop data-driven methods maintaining safety guarantees without requiring precise parameters.

**Operational Challenges** VVC systems address several implementation challenges. Communication delays are managed through state synchronization using prediction models (ARIMA (Guo *et al.*, 2023)) and neighboring bus voltage averaging (Zhang *et al.*, 2023b). Feng *et al.* (2023) focus on rapid voltage recovery during disturbances, while Zhang *et al.* (2023c) examine fast voltage fluctuations from renewables and computational efficiency through DNN projection.

## 6.3 Optimal Power Flow

OPF determines a power system's optimal operating point by minimizing costs while satisfying power flow equations and operational constraints. Several variants address modern power system needs: Security-Constrained OPF (SCOPF) ensures system security under contingencies (Yan and Xu, 2022; Hu *et al.*, 2024), distribution OPF handles voltage regulation with distributed resources (Li and He, 2022), and microgrid OPF manages local resources while maintaining grid interactions (Zhang *et al.*, 2020b; Yu *et al.*, 2024).

These variants share common challenges: AC power flow equations introduce non-convexity making large-scale systems computationally intensive, network constraints must be satisfied at each time step, and energy storage introduces temporal coupling, transforming static OPF into dynamic optimization. Modern power systems add complexity through renewable generation uncertainty requiring probabilistic constraints, fast-changing grid conditions necessitating real-time solutions, and distributed energy resources adding local optimization objectives. Table 6.3 shows how learning-based methods address these challenges, particularly focusing on uncertainty handling and operational constraints while maintaining computational efficiency.

Table 6.3 reveals evolution in problem formulation and solution approaches.[1] Traditional centralized OPF has expanded to distributed

[1] The Problem Setup column presents optimization formulation using state space (S), action space (A), and time resolution (T), with components including Microgrids (MGs), Diesel Generators (DG), Energy Storage Systems (ESS), and problems such as SCOPF and Economic Dispatch (SCED). System Model describes network representations as AC or DC power flow, included components, and uncertainty modeling approaches. Safety Considerations outlines constraint enforcement methods including

**Table 6.3:** Comparison of Safe RL Approaches for OPF and Related Problems

| **Problem Setup** | **System Model** | **Safety Considerations** | **Test Systems** |
|---|---|---|---|
| (Zhang *et al.*, 2020b): Networked MGs; S: load/solar forecast; A: DG/ESS/power transfer; T: 15-min; Distributed multi-agent | AC power flow; Comp: DG/ESS/PV/loads; Uncert: Beta/Gaussian dist; Fixed topology; Smart meter data | Voltage/current limits; DG/ESS/PV limits; Gradient + backtracking; Implicit PF | 33-bus dist. net (5 MGs); Each MG: IEEE 13-bus; Time: 1.4s/agent vs 145.5s central; Vs: DQN, U-PL |
| (Li and He, 2022): Dist. net operation; S: power/storage/price; A: caps/taps/DG/ESS; T: 1h | Net: Black-box; Comp: Full AC dist; Uncert: Ren/load/price; CAISO | Voltage/current/PF limits; Method: CPO; Guarantee: Monotonic; Online safety | IEEE-34/123-node; Data: 2018-20; Vs: DDPG/PPO/SAC |
| (Yan and Xu, 2022): SCOPF; S: P/Q loads; A: V/P control; T: Real-time; Security focus | Net: Full AC; Comp: Gen/loads; Uncert: Load (5%); Steady state | PF/gen/voltage limits; Method: Lagrange-IP; KKT guarantee; RT safe | IEEE 57/300/2000-bus; 99.9% faster; Matches IPOPT quality |
| (Hu *et al.*, 2024): RT-SCED; S: node/gen/SOC; A: 29-dim gen; T: 5-min; Storage focus | Net: DC; Comp: Thermal/wind/PV/ESS; Uncert: Load 2%, ren 5% | Network/gen/ESS limits; Method: Safety layer; Time-coupling safe; 21ms | IEEE 39-bus (30 units); 118-bus test; Vs: PF methods |
| (Yu *et al.*, 2024): Tie-line smooth; S: temp/power; A: chiller flow; T: N/A; Cooling | Net: DC/AC hybrid; Comp: Cooling system; Uncert: PV/cooling load | Temp/flow limits; Method: CVaR; Risk-bounded; Self-adaptive | Zhuhai DCS 144MW; 0.03s solve; Vs: CPO/SAC |

architectures for microgrid coordination, system modeling ranges from classical AC power flow to data-driven approaches, and safety mechanisms have developed from simple penalty methods to constrained optimization frameworks.

### 6.3.1 Safe RL Techniques for OPF and Related Problems

**Constraint-Based Optimization Methods** The core mechanism transforms constrained reinforcement learning into constrained policy search, with direct encoding of system limitations in the optimization objective (see Sec. 2.1.1 for technical details). Li and He (2022) address distribution network operation through CPO, formulating operational constraints for substation capacity, nodal voltages, branch loading, and power factor. Their approach handles network operational limits without penalty coefficient tuning, processes distribution operations as a black box, and accommodates mixed discrete and continuous actions. Zhang *et al.* (2020b) transform the problem into a quadratically constrained

Constrained Policy Optimization (CPO), Model Predictive Control (MPC), Interior Point methods (IP), and Conditional Value at Risk (CVaR), with theoretical guarantees through Karush-Kuhn-Tucker (KKT) conditions or Mixed Integer Programming (MIP). Test Systems describes validation frameworks using IEEE test systems or real systems like District Cooling Systems (DCS), with benchmark comparisons including DDPG, PPO, SAC, DQN, and Unconstrained Policy Learning (U-PL).

linear program, incorporating AC power flow equations to calculate gradient factors and using a backtracking mechanism with coefficient multipliers for marginally violated constraints. Both methods eliminate penalty coefficient tuning through different mathematical frameworks.

**Primal-Dual Methods** Primal-Dual Methods apply Lagrangian relaxation principles to handle constraints in RL (Sec. 2.1.2). Yan and Xu (2022) combine primal-dual deep deterministic policy gradient (PD-DDPG) with classic SCOPF models, incorporating pre-contingency and post-contingency constraints through Lagrangian formulation. Their approach approximates actor gradients by solving Karush-Kuhn-Tucker conditions of the Lagrangian rather than constructing reward and cost critic networks through environmental interactions, enabling faster SCOPF solutions while maintaining contingency handling capabilities.

**Risk-Aware Methods** Risk-Aware Methods incorporate probabilistic risk measures into the reinforcement learning framework. The approach centers on CVaR, which provides a systematic way to measure and control the risk of constraint violations. Risk-aware Soft Actor-Critic (RSAC) (Yu *et al.*, 2024) implements a CVaR-based CMDP formulation to smooth tie-line power fluctuations in grid-connected microgrids with high renewable penetration. Their method employs a district cooling system as a controllable load, incorporating states such as time, temperatures, and power targets while using chiller mass flow rate as the action variable. This approach enables risk quantification and parameter tuning through Gaussian distribution assumptions to address complex thermal dynamics that challenge precise model-based control.

**Safety Filter Architectures** Safety Filter as an RTA architecture involves projection of potentially unsafe actions onto feasible action spaces through optimization-based safety layers. Hu *et al.* (2024) address real-time security constrained economic dispatch (RT-SCED) with energy storage by combining safety exploration (via a safety layer) and safety optimization (via CMDP). Their dual approach projects unsafe actions onto feasible spaces through optimization-based safety layers for single-step constraints while using CMDP formulation for time-coupling constraints, ensuring comprehensive constraint satisfaction.

### Training Approaches

Offline training dominates OPF applications due to operational requirements. Zhang *et al.* (2020b) process historical smart meter data at 15-minute intervals, while Yan and Xu (2022) implement a two-stage process combining supervised initialization with policy refinement. Training incorporates power flow solutions to verify network feasibility, with specific handling of voltage limits and line flow constraints.

Training data sources include historical operational data providing real system behaviors, demonstrated by Li and He (2022)'s use of CAISO market data for distribution optimization, and Monte Carlo simulation generating synthetic scenarios through uncertainty sampling, as in Yan and Xu (2022)'s SCOPF solution with Gaussian-distributed loads. Data resolution directly impacts policy capability to handle temporal dependencies.

OPF exploration mechanisms must maintain both power flow feasibility and operational constraints. Li and He (2022) implement constraint-aware updates through trust region methods in their CPO framework, while projection-based approaches map actions to feasible regions defined by power flow constraints. Hu *et al.* (2024) demonstrate this through a safety layer handling both immediate limits and time-coupling constraints in storage dispatch.

Verification focuses on power system operational metrics, monitoring power flow constraint satisfaction including voltage bounds and thermal limits. Li and He (2022) quantify constraint violations across operating conditions, while Hu *et al.* (2024) verify both immediate and time-coupled constraints for storage operation. Verification addresses normal conditions and contingency scenarios for SCOPF problems.

## 6.3.2 Key Findings and Trends

**Multi-timescale Multi-source Coordination** Temporal coupling from energy storage and renewables requires propagation of safety guarantees across timescales. Safety frameworks must coordinate fast-responding electronic resources with slower mechanical devices while handling time-coupled constraints such as SoC limits (Hu *et al.*, 2024). The interaction between conventional devices and distributed resources introduces coupling in safety constraints (Li and He, 2022), requiring frameworks to handle both resource-specific operational limits and system-level constraints while coordinating storage, distributed generation, and flexible loads (Yu *et al.*, 2024).

**System Security and Constraints** Non-convex power flow equations and N-K security requirements create sparse constraint structures exploitable in RL architecture. Independence of contingency scenarios enables efficient decomposition of safety verification (Yan and Xu, 2022), while safety layers must handle both equality constraints from power flows and inequality constraints from operational limits. Combinatorial security constraints require state space designs capturing essential safety information while remaining computationally tractable, with sparse matrix formulations and independent contingency handling providing mechanisms for scaling safety guarantees (Yan and Xu, 2022).

## 6.4 Demand-Side Management (DSM)

DSM enables grid operators to maintain stability and efficiency by actively controlling energy consumption patterns across building energy management (Khattar and Jin, 2023), district cooling systems (Yu *et al.*, 2023), energy hubs (Qiu *et al.*, 2022), and coordinated storage control (Paesschesoone *et al.*, 2024). DSM applications optimize energy consumption while satisfying operational constraints and maintaining user comfort under uncertainties from renewable generation, user behavior, and market conditions.

Key challenges making DSM suitable for safe RL include balancing competing objectives (cost minimization, emission reduction, comfort maintenance, grid services provision), managing complex dynamics across timescales (from real-time temperature control to day-ahead storage scheduling), and operating under strict constraints (power balance, equipment limitations, user comfort) while handling substantial uncertainties.

Table 6.4 shows recent research approaches varying in problem scope from single buildings to district-level systems and multi-carrier energy hubs. Safety mechanisms differ significantly—some employ explicit safety layers or MPC-based filters guaranteeing constraint satisfaction, while others incorporate safety through guided policy learning or constraint-aware optimization.

Emerging trends in safe RL for DSM include growing focus on multi-agent architectures addressing distributed energy systems (Khattar and Jin, 2023; Zhang *et al.*, 2024a), increasing incorporation of explicit uncertainty handling through prediction modules or robust formulations (Hong and Lee, 2023), and progression toward theoretical safety guarantees during both training and deployment beyond simple penalty-based approaches.

### 6.4.1 Safe RL Techniques for DMS

**Constrained Policy Method** Constrained policy learning directly incorporates safety into policy updates, offering better sample efficiency and scalability. Zhang *et al.* (2024a) develop Consensus Multi-Agent Constrained Policy Optimization (CMACPO), extending CPO to networked multi-agent settings. Their approach achieves system-wide emission targets without centralizing sensitive load information, demonstrating improved convergence compared to traditional distributed optimization approaches and strong potential for large-scale deployment.

**Penalty Method** Hong and Lee (2023) employ a modified DQN with integrated penalty mechanisms for safety violations. Their approach uses theoretically-derived penalty costs ensuring desired safety properties, making soft constraints effectively behave like hard constraints under optimal policy execution. The method exploits short-horizon uncertainty forecasts to achieve robustness and cost-efficiency.

**Optimization-Based Methods** Optimization-based methods use explicit mathematical programming to handle constraints, offering guaranteed feasibility and clear interpretability, though often at higher computational cost.

**Table 6.4:** Comparison of Safe RL Approaches for Demand Side Management

| **Problem Setup** | **System Model** | **Safety Considerations** | **Test Systems** |
|---|---|---|---|
| (Qiu *et al.*, 2022): Multi-energy hub management; S: storage, prices, demand; A: equipment scheduling. Centralized control with hourly timesteps. | CHP, thermal/hydrogen storage, heat pump, and renewables. Explicit modeling of demand and renewable uncertainties. | Safety-guided network adjusts policy to avoid constraint violations. Provides safety guarantees during training and deployment. | UK National Grid dataset with 61 test days. Compared against MILP and LSTM-MPC benchmarks. |
| (Shengren *et al.*, 2023): Distributed energy resource scheduling; S: PV, load, DG generation, and storage SOC. Centralized hourly control for day-ahead scheduling. | Simplified power balance model with DERs (PV, storage, DGs). Quadratic DG cost functions. Handles renewable and load uncertainties. | MIP enforces constraints. Penalty terms in reward during training. Power balance, generation limits, ramping and storage constraints. | One year of demand and PV data. Three DG units and ESS system. Compared against DDPG, TD3, PPO. |
| (Paesschesoone *et al.*, 2024): PV-battery-load system control with states including battery SOC, demand, solar power, and prices. 15-minute control intervals. | Grid-connected system with specified battery characteristics. Models solar, consumption and price uncertainties. | Data-driven MPC safety filter. Battery SOC and power limits enforced. | Validated on Flanders 2021-2022 data. Compared against PPO, TRPO, QR-DQN, DDPG. |
| (Khattar and Jin, 2023): Building energy management with 30-dim state space and 3 continuous actions per building. Decentralized hourly control. | Blackbox building with heat pumps and multiple energy forms. Handles environmental and load uncertainties. | Energy balance, technology constraints, and SOC bounds. Theoretical convergence guarantees of adaptive optimization. | CityLearn Challenge (1st place winner). 4-year simulation compared against SAC, A2C, DDPG, DQN, PPO, TD3. |
| (Hong and Lee, 2023): Energy management with inconsistent supply. States include battery level, demand, generation, prices. Hourly operation. | Energy storage system with defined capacity limits. Models prediction uncertainty as Gaussian random variable. | Theoretical guarantee through (p,q)-failure condition. Penalty cost for failures. Safe RL with short-horizon forecasts. | Korean power company dataset (2017). Compared against DQN, PER, LR-DQN. |
| (Zhang *et al.*, 2024a): Bi-level optimization for low-carbon demand management. Distributed multi-agent control with privacy preservation. | AC power flow with second-order cone relaxation. Models renewable, load, and carbon emission uncertainties. | Generator limits, voltage bounds, line flow limits. Carbon emission constraints. CPO with trust region updates. | Modified IEEE 33-bus and 123-bus systems. Compared against PPO, CPO, MACPO. |
| (Yu *et al.*, 2023): District cooling system control for operating reserve. States include power gap, flow rates, temperatures. Centralized 15-minute control. | Thermodynamic model with chiller, heat exchanger, building components. Occupancy uncertainties. | Safe layer projects unsafe actions using LP. Power, flow rate, and temperature comfort bounds enforced. | Real DCS in Hengqin, China (144 MW). Compared against PI controller, MPC, conventional DDPG. |
| (Hou *et al.*, 2024): Energy storage dispatch; States: Node power levels, ele. prices, SOCs; Actions: Battery charging/discharging decisions; Centralized control | AC power flow distribution system; Uncertain solar generation/demand/ele. prices; Linear battery dynamics | Constraints: Voltage levels, battery charge limits, power flow bounds; Mixed-integer programming integration with DRL | Modified IEEE 34-node network; Implementation: Open-source code available |

Shengren *et al.* (2023) address optimal DER scheduling in renewable-based systems while enforcing operational constraints. Their approach trains a DQN offline to approximate the action-value function $Q(s, a)$ done offline using standard DRL techniques (e.g., experience replay, target network). Once training is complete, the learned Q-network is used for action selection by solving $\max_a Q(s, a)$ subject to constraints on actions, where $Q(s, a)$ is represented as a set of mixed-integer constraints. While ensuring constraint satisfaction, this approach incurs higher computational complexity and depends on Q-function approximation quality.

(Khattar and Jin, 2023) adopts a different approach using adaptive optimization with evolutionary search under trajectory-based guidance. The core technique uses solution functions of optimization (c.f. **??** for an in-depth discussion of this object) as policies while adapting the parameters of the optimization model from online observations. Safety constraints are incorporated through a two-level structure: 1) Upper level: Evolutionary search to optimize policy parameters; and 2) Lower level: Convex optimization to compute actions while respecting operational constraints. The method placed first in the CityLearn Challenge without pre-training (Nagy *et al.*, 2021), demonstrating strong ability to handle multi-building coordination while maintaining individual building constraints.

**Safety Filter Methods** Safety filter approaches separate safety mechanisms from learning, enabling real-time intervention and modularity. Paesschesoone *et al.* (2024) combine MPC filtering with changepoint detection for dynamic power system conditions, implementing continuous model adaptation and policy updates triggered by detected changes. Yu *et al.* (2023) design safety layers for district cooling systems using linear programming to maintain power caps while balancing thermal comfort.Their method's efficient handling of coupled thermal-electric constraints while maintaining real-time performance is notable. Qiu *et al.* (2022) develop LSTM-SDDPG (Long Short-Term Memory-Safe Deep Deterministic Policy Gradient) combining LSTM for uncertainty handling, DDPG for continuous control, and safety-guided networks for constraint satisfaction, effectively managing uncertainty across different energy carriers (electricity, heat, and gas networks) while maintaining operational constraints.

### 6.4.2 Key Findings and Trends

**Multi-resource and Multi-timescale Coordination** The integration of multiple energy carriers has emerged as a key strategy for enhancing system flexibility. Qiu *et al.* (2022) and Khattar and Jin (2023) demonstrate how multi-energy systems leverage different carriers to manage renewable variability effectively. Hong and Lee (2023) address temporal coordination by incorporating both immediate time-of-use pricing and longer-term demand charges, balancing short-term operations with longer-term objectives—essential for practical DSM implementation.

**Balance between Simplicity with Decision-Focused Models** Khattar and Jin (2023) demonstrate that simple predictive models (2-week moving averages) combined with adaptive optimization strategies can be sufficient, challenging assumptions that increasingly complex models are always necessary. Forecasts must be adapted to their downstream decision usage, finding appropriate complexity levels ensuring reliable operation while maintaining implementability.

**Operational Safety and Constraint Satisfaction** Safety constraints are fundamental for real-world adoption (Shengren *et al.*, 2023), particularly challenging in nonstationary environments where changepoint detection may be employed (Paesschesoone *et al.*, 2024). Yu *et al.* (2023) introduce specialized self-adaptive methods managing recovery processes and preventing rebound peaks in district cooling systems, while Zhang *et al.* (2024a) extend safety considerations to include environmental impacts through nodal carbon intensity management.

## 6.5 Critical Load Restoration

CLR strategically re-energizes critical loads following outages while maintaining operational constraints and maximizing system resilience. Modern distribution networks face increased complexity due to DER integration, multiple microgrids, complex ownership structures, extreme weather events, and intermittent renewable resources, making efficient CLR strategies essential for grid reliability and minimizing outage impacts.

CLR encompasses post-fault service restoration, microgrid formation, DER dispatch optimization, and network reconfiguration. The process must address cold load pickup effects, load prioritization, DER uncertainty, and dynamic resource availability. Networked microgrids with mixed ownership environments further complicate the problem by necessitating distributed decision-making frameworks.

Key technical challenges in CLR include: (1) the combinatorial nature of switching operations leading to exponential growth in the action space, (2) the need for fast decision-making under uncertainty, particularly with renewable DER integration, (3) maintaining system stability and operational constraints during the restoration process, and (4) coordinating multiple resources and stakeholders while ensuring both local and system-wide objectives are met.

Table 6.5 summarizes recent approaches. Action space complexity solutions include action masking (Vu *et al.*, 2023), end-to-end acceleration (Wang *et al.*, 2023d), and efficient binary switching representations (Jacob *et al.*, 2024). DER uncertainty handling employs memory-enhanced design with capacity predictions (Fan *et al.*, 2024), meta-learning (Abdeen *et al.*, 2024), and explicit probabilistic models (Si *et al.*, 2024). System stability is maintained through episode termination and reward shaping, while multi-resource coordination uses distributed multiagent control (Vu *et al.*, 2023; Si *et al.*, 2024).

Methods vary in model detail from linearized approximations to detailed three-phase unbalanced representations, reflecting trade-offs between computational efficiency and accuracy. Unique considerations include cold load pickup effects (Wang *et al.*, 2023d), dynamic microgrid boundaries (Si *et al.*, 2024), and restoration sequence constraints (Li and Wu, 2024). Successful CLR implementations require balancing modeling fidelity, computational tractability, and practical operational requirements.

### 6.5.1 Safe RL Techniques for CLR

**Reward Shaping Methods** Reward shaping approaches modify the reward function to incorporate safety penalty without enforcing hard constraints. Du and Wu (2022) employ a two-stage framework using expert demonstrations for safe initial policies before online learning. Their method enforces DG and ES limits through action clipping while managing power balance and voltage constraints via reward shaping. Jacob *et al.* (2024) introduce a structured reward combining total energy

**Table 6.5:** Comparison of Approaches for Critical Load Restoration

| **Problem Setup** | **System Model** | **Safety Aspects** | **Test Systems** |
|---|---|---|---|
| (Du and Wu, 2022): Service restoration in islanded microgrids. Expert demonstrations for pre-training. | Linearized DistFlow equations. DER uncertainties. | Voltage/power flow constraints. Action clipping, reward shaping. | IEEE 123-node. ERCOT load data. Hourly power imbalance percentage. |
| (Wang *et al.*, 2023d): Sequential restoration with cold load pickup. Binary component status. | Linearized power flow. Multi-period. Full observability. | Flow, voltage, generator limits. Radiality. Good-Turing bounds. | IEEE 33/123-bus, 1069-bus. vs. Gurobi. |
| (Vu *et al.*, 2023): Priority-weighted load restoration in networked microgrids. Distributed multi-agent control for mixed ownership environments. | OpenDSS-based power flow. Networked microgrid topology. Cold load pickup effects. | Power balance, voltage limits, generator bounds via invalid action masking. | IEEE 13/123/8500-node systems. |
| (Fan *et al.*, 2024): POMDP max. weighted load restoration. A: breaker states, restoration levels. | Three-phase unbalanced model. DER uncertainties. | Power flow, voltage bounds. Episode termination for constraint breach. | IEEE 123-bus (15 nodes). 7-day DER forecasts. vs. DQN, DDPG, SAC. |
| (Abdeen *et al.*, 2024): MDP with renewable forecasts, load levels, battery SOC. Actions control DER outputs. | Three-phase model. Stochastic renewables. Constant load during outage. | Voltage limits. DER bounds on storage, fuel. Regret bounds. | IEEE-13 bus (15 loads, 4 DERs). 27 scenarios. vs. warm-start and ES-RL. |
| (Jacob *et al.*, 2024): MDP for reconfiguration and load shedding. S: node/edge variables, topology. Binary switching actions. | Three-phase unbalanced model. Grid-forming/feeding DERs. Rand. outages. | Voltage/power flow limits. Switch masking. | IEEE 13/34/123-bus. ms-scale tests. vs. MISOCP. Energy served, voltage regulation. |
| (Si *et al.*, 2024): POSG with PV, load, switch agents. S: voltage, line loading, switch status. CTDE. | Power flow with radial topology. PV uncertainty. Single switch per step. | Action masking (power, topology). Reward shaping. | IEEE 123-node. vs. VDN and model-driven OPT. |
| (Khattar *et al.*, 2025): CLR under uncertain topology changes. Hierarchical structure with cell-level and coordinating agents. | Distribution grid with dynamic topology. CTDE approach. | Topology-dependent action masks. Power flow and operational constraints. | IEEE 123-bus with tie/sectionalizer switches. Topology contingency scenarios. |

supplied to loads ($E_{supp}$) with voltage violation penalties ($V_{viol}$) for three-phase measurements across all buses, implemented as $r(s,a) = E_{supp} - V_{viol}$ when power flow converges, and zero otherwise. Graph Capsule neural networks are used to capture node properties and edge information for network reconfiguration decisions. Li and Wu (2024) enhance DQN with Artificial Potential Fields addressing sparse reward problems, while Abdeen *et al.* (2024) develop First-Order Meta-based RL with Evolution-Strategy RL to avoid computationally expensive second-order derivatives. Fan *et al.* (2024) combine spatial and temporal features in their Recurrent Graph Soft Actor-Critic, using episode termination for severe violations alongside reward penalties for minor constraint breaches.

**Safety Filter Methods** Wang *et al.* (2023c) introduce MT-PIPPO (Multi-Task Physics-Informed Proximal Policy Optimization) combining PPO-based MARL, multi-task learning for different network topologies, and two-stage safety verification. Their approach checks energy dispatch feasibility using physics-informed constraints, then applies correction optimization to find the nearest feasible action when violations occur. Wang *et al.* (2023d) develop an end-to-end framework combining deep learning with MILP, where ML predicts binary variables and binding constraints to transform complex MILP into simpler LP problems while maintaining safety through explicit optimization constraints.

**Action Masking** Action masking enforces safety constraints by explicitly preventing unsafe actions before they can be selected (Si *et al.*, 2024; Vu *et al.*, 2023). The technique introduces a binary mask $m_t \in 0, 1^{|A|}$ at each timestep $t$, where $m_t[a] = 0$ indicates invalid actions and $m_t[a] = 1$ denotes valid actions. The masked Q-values are computed as $Q_{\text{masked}}(s_t, a) = Q(s_t, a) \cdot m_t[a] - M \cdot (1 - m_t[a])$, where $M$ is a large negative value (typically 1e8). This implementation ensures invalid actions have negligible selection probability during both exploitation and exploration phases, effectively reducing the action space from $2^{|A|}$ to only valid actions. For multi-agent settings (Vu *et al.*, 2023), when agents propose joint actions, the system checks for constraint violations. If violations occur, a random agent is selected to modify its action, with its highest Q-value action masked. This process continues until a valid joint action set is identified. The action masking technique demonstrates significant scalability, with (Vu *et al.*, 2023) showing successful implementation on systems ranging from 13 to 8500 nodes. The method improves learning efficiency by constraining the feasible action space and preventing exploration of invalid actions, leading to faster convergence to feasible strategies.

**Multi-Agent Methods** MARL addresses power network spatial distribution by treating network components as individual agents, combining with reward shaping (Fan *et al.*, 2023a; Si *et al.*, 2024), safety filters (Wang *et al.*, 2023c), and action masking (Si *et al.*, 2024; Vu *et al.*, 2023; Khattar *et al.*, 2025) to ensure safety requirements.

Fan *et al.* (2023a) implement graph-based attention where agents coordinate through self-attention while maintaining individual decision capabilities, using CTDE and graph convolutional networks that balance local information with global network topology. Si *et al.* (2024) develop Dynamic Agent Network architecture with QMIX allowing arbitrary-sized neighboring agent coordination, where attention mechanisms learn importance weights for interactive agents—particularly for dynamic microgrid boundaries.

Wang *et al.* (2023c) combine multi-agent coordination with physics-informed safety verification through MT-PIPPO, using Dec-POMDP framework with local observation spaces and shared reward functions alongside physics-based safety filters. Vu *et al.* (2023) integrate action masking with multi-agent Deep Q-Learning, sharing invalid action information while using OpenDSS simulation for constraint verification. Khattar *et al.* (2025) address uncertain topology changes through hierarchical MARL that divides distribution grids into cells with independent control agents and a coordinator for inter-cell power transfer, demonstrating superior generalization to unseen structural disruptions in IEEE-123 bus system experiments. A key innovation is their topology-dependent action masks mechanism, which dynamically identifies unavailable actions after topology changes, addressing a limitation in existing methods that assume fixed topology during restoration.

### 6.5.2 Key Findings and Practical Considerations

**Sequential Decision-Making and Temporal Dependencies** CLR requires coordination between immediate switching actions and longer-term power dispatch—a multi-timescale challenge fundamentally different from single-timestep demand-side management. Fan *et al.* (2024) implement a 4-hour restoration horizon discretized into 16 steps, while Li and Wu (2024) demonstrate that generator start-up sequences must optimize both temporal and power resource allocation rather than following shortest paths. The process involves hierarchical dependencies between self-starting black-start generators and non-black-start generators requiring external power, with effective strategies considering parallel restoration paths and resource utilization.

Wang *et al.* (2023d) also recognize the Cold Load Pickup (CLPU) effects where restored loads temporarily demand more power than their steady-state values due to loss of diversity in thermostatically controlled loads. This is modeled through $P_{j,t}^{dem} = l_{j,t}\rho_{j,t}^{clpu}P_{j,t}$, where $P_{j,t}^{dem}$ is the actual load demand at bus $j$ and time $t$, $l_{j,t}$ is the load energization indicator (1 if load is energized at time $t$), $\rho_{j,t}^{clpu}$ is the CLPU ratio representing the magnitude of load spike (>1), and $P_{j,t}$ is the normal steady-state load demand. This effect necessitates sequential restoration planning to prevent system overload, creating temporal coupling in the optimization through the restoration periods.

**Multi-Energy Integration** Multi-energy systems operate on different timescales, with rapid power flow changes contrasting slower gas and heat network dynamics (Wang *et al.*, 2023c). This requires careful control timestep selection and physics-informed safety layers handling these coupled but different-speed dynamics. Constraints in one energy vector directly impact others, requiring coordination between different DER types including distributed generators, energy storage, and PV systems (Du and Wu, 2022). Systems must also account for hierarchical relationships between public-managed and private-managed DERs, as operators can only directly control public resources (Fan *et al.*, 2023a).

**Partial vs Complete Blackouts** During partial blackouts, operational generators in “islanded” systems can supply power to outage areas, enabling different restoration strategies compared to complete blackouts (Li and Wu, 2024). Black-start DGs play a critical role in maintaining voltage and frequency stability, with DG output constraints requiring immediate episode termination if violated (Si *et al.*, 2024).

**Network Topology** Distribution system restoration fundamentally requires handling dynamic microgrid boundaries (Si *et al.*, 2024; Khattar *et al.*, 2025). Network reconfiguration affects power flow patterns and system stability, making it a graph manipulation problem rather than just load control (Jacob *et al.*, 2024). Furthermore, distribution systems require explicit radiality constraints (Wang *et al.*, 2023d). During extreme events, most network structure typically remains intact, enabling multi-task learning frameworks where different network topologies are treated as related but distinct tasks (Wang *et al.*, 2023c). Graph neural networks effectively capture the combinatorial nature of switching decisions and physical network connectivity (Jacob *et al.*, 2024; Fan *et al.*, 2024).

**System Integration** Integration occurs through direct control of dispatchable generators and energy storage systems, existing DER monitoring systems, and SCADA systems for measurement and control (Fan *et al.*, 2023a). A key aspect is to position systems as decision support tools rather than autonomous controllers (Li and Wu, 2024). The method connects with outage management systems for fault message collection and transmission to multi-class classifiers (Wang *et al.*, 2023d). Mixed ownership environments, where microgrids belong to different utility or non-utility owners, make centralized control impractical and necessitate distributed control architectures (Vu *et al.*, 2023).

## 6.6 EV Charging and Coordination

Rapid EV adoption introduces significant power demands that can strain grid infrastructure while offering potential grid services through V2G capabilities and demand flexibility. The EV charging coordination problem spans multiple operational levels: distribution system operators must manage charging schedules across stations while respecting network constraints; charging station operators must allocate power among EVs with heterogeneous requirements; and route planning must optimize paths considering energy consumption and charging availability.

Operational challenges include component-level constraints (battery SoC limits, charging rate bounds, powertrain limits), system-level constraints (power balance, station capacity, grid stability), and significant uncertainties in user behavior, electricity prices, and renewable generation. Table 6.6 summarizes recent safe RL approaches addressing these challenges across various problem formulations and safety mechanisms.

### 6.6.1 Safe RL Techniques for EV Charging and Coordination

**Constrained Policy Optimization (CPO)** Li *et al.* (2019) adapt CPO for EV charging by integrating charging/discharging constraints through a cost function handling battery SoC limits. Their method employs trust regions enabling stable policy updates while maintaining charging demand satisfaction.

**SAC-based Extensions** Zhang *et al.* (2023d) develop Constrained SAC incorporating constraint handling through Lagrangian relaxation and rule-based safety filters. Their approach uses multiple constraint terms for battery capacity limits, charging requirements, and final state constraints, while introducing a novel EV grouping strategy (waiting, eligible for V2G, charging, fully charged) reducing action space dimensionality. Yang *et al.* (2025) introduce Augmented Lagrangian SAC addressing standard Lagrangian limitations through quadratic penalty terms, incorporating real-time electricity prices and charging patterns into constraint formulations.

**Penalty-based Methods** Zhang *et al.* (2023f) implement multiple reward components including penalties for battery depletion and inefficient charging, with episode termination when battery levels become unsafe. Jiang *et al.* (2021) use penalty functions during training for insufficient/excessive charging, implementing hard constraints through calibration during deployment. Biswas *et al.* (2024) develop Physics-informed Exploration maintaining feasible action ranges, tracking infeasibility through counters and updating action bounds based on observed violations.

**Table 6.6:** Summary of Safe RL Methods for EV Applications

| Problem Setup | System Model | Safety Aspects | Test Systems |
|---|---|---|---|
| (Li *et al.*, 2019): EV charging scheduling. CMDP w/ cont. charging actions. | Battery charging dynamics with energy loss model. Price-taker. | Battery energy limits, charging rate constraints. CPO. | MISO electricity price data (2017-2018). vs. DQN, DDPG, Safety-Prissy, MPC. |
| (Jiang *et al.*, 2021): Coordinated charging in parking lot. Min. load variance s. t. energy demands. | Maximum charging rate per bay. Statistical patterns in vehicle dynamics. | Energy demand satisfaction, rate limits. Action calibration for constraint handling. | 15-20 charging bays, 52-week training data. PyTorch implementation. |
| (Zhang *et al.*, 2020a): Plug-in Hybrid EV energy management. CMDP with neural network strategy actor. | Power-split hybrid with double planetary gear. Battery internal resistance model. | Component speed/ power/torque limits, battery bounds. Coach intervention for safety. | Dublin Bus route data. Cloud-based training with onboard deployment. |
| (Zhang *et al.*, 2023f): EV route planning with charging. MDP using power and loops state space. Two-layer hierarchical control. | Graph w/ distance/energy costs. Deterministic travel times, piecewise charging. | Battery capacity and energy feasibility constraints. Early termination for low battery. | Sioux Falls (24 nodes) and Beijing network (83,917 nodes). Comp. w/ Integer Lin. Program., Dijkstra, A*. |
| (Biswas *et al.*, 2024): Hybrid EV mgmt. State: power, SOC, velocity. Cont. power actions. | Toyota Hybrid System powertrain. Component efficiency maps. | Power, SOC, current constraints. Uses safety layer and physics-informed exploration. | Urban Dynamometer Driving Schedule training, Artemis Urban Driving Cycle testing. |
| (Yang *et al.*, 2025): EV charg. scheduling under uncertainty. CMDP with continuous actions. | Simplified charging model focusing on SOC dynamics. Price-taker. | SOC limits, charging rate constraints. Augmented Lagrangian method. | German electricity market data (2018-2020). Compared with DDPG, SAC, MPC. |
| (Zhang *et al.*, 2023d): Microgrid profit maximization. CMDP with states for time, photovoltaic generation, load, battery. | Microgrid with photovoltaic system, storage, loads. Nonlinear charging with Vehicle-to-Grid capability. | Power balance, storage limits, charging constraints. Constrained SAC with safety filter. | 100 charging piles, 3 vehicle types, 600kWh battery. vs. CPO, Lagrangian SAC. |
| (Zhang *et al.*, 2024d): Joint charg. and computat. allocation. Hierarchical two-timescale control. | SOC dynamics model. Uncertain vehicle arrivals and tasks. | SOC limits, power constraints, task deadlines. Lyapunov safety guarantees. | 25 charging piles, 4 vehicle models. 24% load variance improvement. |
| (Zhang *et al.*, 2023a): Real-time charging with grid commands. Two-stage Distributed PPO. | Uncertain travel patterns. | SOC bounds, charging safety module, bias elimination. | 20 vehicles. vs. Twin Delayed DDPG, SAC, PPO. |

**Runtime Assurance Architectures** Zhang *et al.* (2020a) implement a Simplex architecture with coach-actor-double-critic framework, where the coach provides a rule-based CD-CS (Charge-Depleting-Charge-Sustaining) fallback strategy while dual variables handle constraint satisfaction. When actions exceed feasible ranges, the coach intervenes with penalty additions to help the actor learn feasible actions. Zhang *et al.* (2023a) introduce two safety modules: an EV Charging Safety Module modifying charging/discharging power based on current SOC, and an Allocation Bias Elimination Module employing Advantage Least Laxity First (ALLF) using advantage functions to quantify charging flexibility beyond immediate time constraints. Zhang *et al.* (2024d) develop a safety-filter architecture with Lyapunov Constraints transforming long-term charging constraints into state-wise conditions defining action safety, and a State-Cost Action Function generating actions satisfying these constraints. This approach transforms a constrained QP problem into an unconstrained one where candidate safe actions are preselected by the State-Cost Action Function.

### Training Approaches

Various approaches employ offline training with simulated environments. Li *et al.* (2019) sample 500 trajectories per iteration using simulated charging scenarios. Zhang *et al.* (2020a) implement cloud server training using historical driving cycles (80% training, 20% testing). Yang *et al.* (2025) employ off-policy training with German day-ahead market price data, enabling historical experience reuse without online interaction.

Safe exploration strategies include Zhang *et al.* (2020a)'s $\epsilon$-greedy annealing with coach intervention verification, Biswas *et al.* (2024)'s state-dependent action ranges updated based on encountered infeasibilities, and Yang *et al.* (2025)'s maximum entropy regularization with double-critics networks avoiding overestimation bias.

Data requirements vary across methods. Jiang *et al.* (2021) generate 52-week dynamic EV arrival/departure traces for training with 4-week validation/testing traces. Zhang *et al.* (2023d) utilize PV generation patterns across four weather types with Gaussian-distributed load profiles. Zhang *et al.* (2023a) incorporate both real-world charging data and simulated grid conditions for their two-layer safety filter.

## 6.6.2 Key Findings and Considerations

**Model and User Behavior Aspects** Zhang *et al.* (2023d) and Zhang *et al.* (2023a) incorporate nonlinear EV charging characteristics in battery models, accounting for varying charging rates during constant-voltage stages. For user behavior, Zhang *et al.* (2023d), Yang *et al.* (2025), Jiang *et al.* (2021), and Li *et al.* (2019) implement Gaussian-distributed initial SOC and arrival times calibrated to commuting patterns, e.g., CBD rush hours. Parking duration models vary from Gaussian distributions (Jiang *et al.* (2021) using mean 8h, std 1.5h) to arrival-departure time differences.

**V2G Integration** Zhang *et al.* (2023d) and Zhang *et al.* (2023a) define V2G operation windows aligned with grid demands, limiting V2G operation below 0.25 SoC to minimize battery degradation. Yang *et al.* (2025) and Zhang *et al.* (2023a) model bidirectional power flow with continuous charging/discharging rates. Zhang *et al.* (2023a) quantify flexibility contribution considering both immediate and long-term impacts of charging schedule deviations, while Jiang *et al.* (2021) address uncoordinated charging through load profile characterization.

For multi-stakeholder considerations, Zhang *et al.* (2023d) and Zhang *et al.* (2024d) balance grid operator profits, EV user requirements, and system-level objectives. Zhang *et al.* (2024d) implement pricing mechanisms linking charging prices to computation contribution, creating relationships between grid services and edge computing resources.

**Multi-timescale Operation** Zhang *et al.* (2023d) and Zhang *et al.* (2023a) implement multi-level control architectures, with Zhang *et al.* (2023a) using DSO/Charging Stations/EVs hierarchy addressing large-scale EV control scalability. Zhang *et al.* (2020a) demonstrate price-based coordination between electricity and fuel consumption through hierarchical Onboard Units for real-time control with cloud-based training systems. Control intervals range from 15-minute periods to daily optimization horizons.

**Integration Considerations** Zhang *et al.* (2023a) address FERC Order No. 2222 requirements for DER aggregation, noting distribution factor determination challenges. Li *et al.* (2019), Jiang *et al.* (2021), Zhang *et al.* (2023a), and Yang *et al.* (2025) outline infrastructure modifications including continuous rate control capability, real-time SOC monitoring, and charging rate control systems.

## 6.7 Power System State Estimation

Traditional power system state estimation relies on model-based algorithms (e.g. weighted least squares for static state estimation or Kalman filters for dynamic estimation). Recent advancements explore RL for data-driven enhancement, particularly under complex conditions.

RL techniques have been applied to improve state estimation accuracy in both static and dynamic contexts. For example, Yuan *et al.* (2022) develop a hierarchical deep actor-critic RL framework that treats distribution system state estimation as a sequential decision problem, jointly estimating system states and optimizing measurement selection in real-time. This model-free approach demonstrates superior performance in unobservable conditions compared to traditional weighted least squares methods. On the dynamic estimation front, Hu *et al.* (2020) provide theoretical guarantees of estimation error convergence for nonlinear systems, integrating Lyapunov stability theory with deep RL to ensure convergence even under model uncertainties and missing data. Zhang *et al.* (2024c) use Deep RL for adaptive forecasting-aided state estimation in distribution networks, outperforming Kalman filters with hybrid measurements.

MARL facilitates distributed state estimation across control areas, substations, or devices. Salamat *et al.* (2023) introduce Distributed RL State Estimation where sensor nodes use local RL estimators with consensus filters for global coherence, achieving faster tracking without prior dynamic models. These approaches align with distributed energy resources and PMU networks, offering scalability and resilience to topology changes or failures.

Hybrid techniques blend physical modeling with data-driven learning. Liu *et al.* (2024a) develop physics-inspired neural networks for secondary distribution networks using power flow equation structures to guide network architecture, constraining models to obey Kirchhoff's laws while requiring less training data than black-box approaches. Similarly, Habib *et al.* (2023) propose Deep Statistical Solver using graph convolution to propagate information along feeder topology with weak supervision from power flow equations, capturing complex spatial correlations by treating grids as graphs to improve robustness against bad or missing data.

**Future Research Directions** RL offers model-free adaptation, handles nonlinearities, provides computational efficiency post-training, and enables active decision-making. Achieving theoretical convergence guarantees through adaptive control and Lyapunov stability analysis is essential, while integrating graph-based insights can enhance robustness, as demonstrated by Jin *et al.* (2020)'s boundary defense mechanism leveraging network topology to isolate attacked regions. Future methods should combine physics-informed robustness with data-driven adaptability for superior performance. Future research should prioritize safe RL estimators with bounded errors, rapid bad data rejection (e.g., by using federated learning to improve cybersecurity while preserving data privacy (Kesici *et al.*, 2024)), and robust contingency reconfiguration. For MARL, exploring distributed reward shaping, consensus-based updates, and graph neural network critics for multi-area estimation is crucial. Achieving theoretical convergence guarantees through adaptive control and Lyapunov stability analysis is essential, while integrating graph-based insights can enhance robustness, as demonstrated by Jin *et al.* (2020)'s boundary defense mechanism leveraging network topology to isolate attacked regions. Future methods should combine physics-informed robustness with data-driven adaptability for superior performance.

## 6.8 Cybersecurity in Power Systems

Modern power systems face sophisticated cyber threats targeting communication networks and control devices. Conventional security measures often struggle with evolving attacks, prompting growing interest in RL for autonomous and adaptive defense.

**Detection Systems** RL enhances both intrusion and anomaly detection by enabling adaptive learning without manual reconfiguration. The goal is typically to identify unusual or suspicious deviations in system measurements, control signals, or operational patterns that may indicate cyberattacks (e.g. falsified data, malicious control commands) or incipient failures. For example, Kurt *et al.* (2018) pioneered online cyber-attack detection using POMDP formulation with a model-free algorithm that trains defenders with low-magnitude attacks, employs sliding observation windows, and demonstrates superior performance across various attack scenarios including false data injection, jamming, denial of service, and network topology attacks. (Hu *et al.*, 2022a) proposed RL-driven Adaptive Feature Boosting, where agents dynamically adjust focus on different data features, achieving approximately 97.3% detection accuracy—a 5.5% improvement over non-RL baselines.

RL has proven particularly effective for detecting False Data Injection (FDI) attacks. Chen *et al.* (2018) model FDI attacks as a POMDP, using Q-learning and kernel-density-based detection to adaptively identify malicious data manipulations. Recent approaches leverage federated learning, enabling operators to jointly build detection models without sharing private measurements (Kesici *et al.*, 2024). Additionally, Gautam (2023) employ RL for optimal PMU placement, enhancing system-wide observability and robustness against FDI attacks. Beikbabaei *et al.* (2025) develops model-free approaches protecting both grid-forming and grid-following inverters without requiring internal control parameters.

**Secure Communication** Agents observe channel conditions and jammer behavior to learn optimal spectrum usage strategies, significantly improving network reliability under attack. RL bolsters communication security through dynamic adaptation of network configurations. Anti-jamming applications optimize frequency hopping, power control, and routing to overcome attacks on wireless sensor networks (Luo *et al.*, 2022). Xu *et al.* (2022) apply deep RL-based moving-target defense randomizing data paths in IP networks to confuse adversaries, reducing attackers' ability to predict communication paths.

**Threat Mitigation and Resilient Control** RL plays a crucial role in active threat mitigation, orchestrating control actions to maintain power system stability during attacks. Maiti and Dey (2024) propose safe deep RL frameworks where agents trigger protective actions (relay tripping, load shedding, reconfiguration) with formal verification ensuring the controller will not drive the system into unsafe conditions. Resource-constrained defense approaches use RL with temporal logic specifications to prioritize actions (Moradi *et al.*, 2023). Agents satisfy high-level goals (encoded as logical formulas) while finding optimal compromises between competing objectives like sustaining service versus isolating compromised areas. Zhang *et al.* (2024b) use RL to mitigate cascading failures, generating adaptive load shedding strategies in real-time as attacks unfold. Safe exploration techniques ensure agents do not inadvertently destabilize the system during training or execution.

**Multi-Agent Approaches** MARL features prominently across all cybersecurity applications. For detection, decentralized agents at different network nodes learn to detect attacks in their vicinity, with attention mechanisms prioritizing critical alerts (Sethi *et al.*, 2021). Such distributed systems scale better with grid size and heterogeneity, achieving higher detection rates with lower false alarms by combining local observations. Adversarial approaches frame detection as a game between defender and attacker agents, creating more robust systems capable of catching sophisticated adversaries (Mouyart *et al.*, 2023).

For communication security, distributed MARL-based anti-jamming algorithms enable each wireless node to sense the spectrum and choose communication channels. Through collaborative learning, nodes coordinate to avoid jammed frequencies and establish resilient links (Ma *et al.*, 2024).

In threat mitigation, heterogeneous MARL frameworks employ specialized agents for different control actions (line reconfiguration, generator redispatch, load shedding), with coordinators ensuring optimal combined responses (Moradi *et al.*, 2024). This approach allows systems to survive complex attack scenarios—isolating compromised substations while rerouting power flow to prevent overloads. For networked microgrids with coupled dynamics, Mukherjee *et al.* (2024) demonstrate that a vertical variant of federated reinforcement learning outperforms fully decentralized architectures by enabling privacy-preserving parameter sharing between agents while capturing system-wide electrical interactions.

**Future Research Directions** By learning optimal responses to complex attack sequences, RL agents help grids absorb and recover from attacks with minimal disruption, adapting to new tactics while coordinating across networks. Future research should focus on enhanced safety verification through formal methods, scalable MARL frameworks coordinating across hierarchical control layers, transfer learning for adapting security policies across different grid topologies, integration of operator knowledge into training, and adversarial training for anticipating emerging attack vectors.

# 7

# Simulation Environments and Benchmarks

The preceding chapters established how safe RL can address power system control challenges by tailoring algorithms (Chapters 2–5), architectures (Chapter 5), and application-focused formulations (Chapter 6). However, the proprietary nature of industrial software and the wide-ranging simulation needs in power systems have spurred the development of open-source, domain-specific environments. These tools not only enable realistic training and evaluation of RL-based strategies but also promote reproducibility and fair comparisons across different methods. Table 7.1 offers a high-level comparison of the leading open-source simulation frameworks for power systems. Each entry summarizes core features (e.g., voltage regulation, EV coordination), typical approaches to safety constraints, RL integration compatibility, and reference benchmark methods.

**EV Charging and V2G** ACN-Sim (Lee *et al.*, 2021) is a widely used simulator for EV charging coordination, focusing on realistic infrastructure constraints such as transformer and line current limits, J1772 compliance, and unbalanced three-phase systems. It integrates real charging data (ACN-Data), ties into multiple power system tools (MAT-

**Table 7.1:** Comparison of Open-Source RL Simulation Environments for Power Systems

| **Overview** | **Safety & MARL** | **Integration & Model** | **Evaluation** |
|---|---|---|---|
| (Vazquez-Canteli *et al.*, 2020) CityLearn: District-level DSM; Standardized environment; No co-simulation requirement | Thermal demands; Storage limits; Centralized/Decentralized; Blackbox model | OpenAI Gym; Interfaces with standard data formats (building data/characteristics, weather); SAM for PV | Rule-based controller, SAC; Multiple building types/climate zones; CityLearn Challenge with multiple metrics |
| (Lee *et al.*, 2021) ACN-Sim: EV charging coordination; Infrastructure planning; Modular, object-oriented architecture | Infrastructure constraints (e.g., J1772 standards, transformer limits); Centralized control | OpenAI Gym; MATPOWER; PandaPower; OpenDSS; ACN-Data for real data and ACN-Live for field testing | 12,000+ unit/integration tests; Baselines( e.g., Round Robin, Least-Laxity First, MPC) |
| (Pigott *et al.*, 2022) GridLearn: VVC via building management (smart inverters, storage, flexible loads); Grid-/building-level obj. | Hard constr.: thermal deadbands, voltage limits, power factor; 96 independent agents; Decentralized MARL | CityLearn; OpenAI Gym; PettingZoo; Stable Baselines; Multiple climate zones/building models | Rule-based control baseline; IEEE 33-bus network |
| (Biagioni *et al.*, 2022) PowerGridWorld: Building coordination for VVC; Heterogeneous system control (building, PV, EV) | Soft voltage limits and thermal comfort; Cooperative/competitive MARL; Heterogeneous agents | OpenDSS; EnergyPlus; RLLib; CityLearn/GridLearn | MADDPG; PPO; Homogeneous/heterogeneous agent scenarios |
| (Henry and Ernst, 2021) Gym-ANM: DSM; Min. loss via DERs and generator control | Voltage/current/SOC limits; Penalty-based constraints | OpenAI Gym; Stable-Baselines3; Power flow models; Newton-Raphson solver | MPC, PPO, SAC; Multiple scenarios (windy night, high EV demand, high renewables) |
| (Fan *et al.*, 2022) PowerGym: VVC with device coordination (voltage regulators, switchable capacitors, and batteries) | Hard constr.: voltage bounds, device limits, SOC constraints; Centralized | OpenDSS; OpenAI Gym; Stochastic load profiles; Multi-phase modeling | PPO, SAC; IEEE systems (13/34/123/8500-bus) |
| (Orfanoudakis *et al.*, 2024) EV2Gym: EV charging optimization; Comprehensive V2G modeling | Transformer/battery/station constraints (normalization or penalty); Cooperative MARL | SB3; Real EV data; Market pricing; PV generation | Heuristics e.g., Round Robin, MPC, baselines from SB3; Tested up to 10K charging stations |
| (Sahu *et al.*, 2023) DSS-SimPy-RL: CLR via network reconfig.; Network rerouting against congestion and cyber threats | Voltage/queue limits; Channel capacity; Resilience metrics; Centralized | OpenDSS; SimPy; PowerGym for VVC; SB3; Cyber-physical modeling | Spanning tree; DQN/PPO/A2C; Expert heuristics; N-1/2/3 contingencies; IEEE systems (13/34/123/8500-bus) |
| (Yeh *et al.*, 2024) SustainGym: EV charging; Market bidding; Data center scheduling; Cogeneration; Building control | Physical constraints via penalties; No hard constraints; MARL environments (EV, Cogen, Building) | ACNSim; IEEE RTS-GMLC; EnergyPlus; RLLib; SB3; Distribution shifts (demand, environment) | SAC, PPO, MPC; Carbon emissions; Distribution shift evaluation |

POWER, PandaPower, OpenDSS), and wraps seamlessly with OpenAI Gym for RL-based scheduling or load-flattening tasks. With thousands of built-in tests and a field testing platform (ACN-Live), ACN-Sim provides a practical baseline for comparing new algorithms against well-established approaches.

EV2Gym (Orfanoudakis *et al.*, 2024) extends the focus to V2G scenarios. It embeds battery degradation models and scales up to thousands of charging stations, a feature particularly relevant for large-scale load management or aggregator-based scheduling. Real EV data (ElaadNL, RVO-NL), standardized Gym interfaces, and support for multiple control strategies (RL, mathematical programming, heuristics) make EV2Gym a flexible test bed for profit maximization and setpoint tracking tasks under transformer and battery constraints.

**Building Energy Management and Demand Response** CityLearn (Vazquez-Canteli *et al.*, 2020) provides a standardized OpenAI Gym environment for building energy management and demand response. Its core tasks include load shaping, coordinated thermal and electrical storage, and peak demand reduction across multiple buildings. CityLearn leverages pre-simulated data rather than full co-simulation, balancing computational tractability with realism. Constraints (comfort, storage bounds) are enforced by action clipping. The environment includes both centralized and multi-agent modes, with a rule-based controller (RBC) and a reference SAC implementation as baselines. This platform is used for CityLearn Challenge with results compared on metrics such as ramping, load factor, and peak demand.

PowerGridworld (Biagioni *et al.*, 2022) centers on building coordination for voltage regulation and heterogeneous resource management in multi-agent setups. Its modular architecture supports both homogeneous and heterogeneous agents, allowing flexible combinations of buildings, PV systems, and EV chargers under cooperative or competitive settings. Constraints (e.g., voltage limits, thermal comfort) are handled via soft penalty terms in the reward function rather than hard safety checks. It integrates with OpenDSS for power flow calculations, EnergyPlus for building modeling, and RLLib for RL algorithm implementations (e.g., MADDPG, PPO).

**Distribution Network Management (VVC, actine network management)** GridLearn (Pigott *et al.*, 2022) tackles voltage regulation challenges in PV-rich grids by coordinating distributed resources such as smart inverters and flexible loads. Its key feature is the decentralized multi-agent approach (up to 96 agents) that focuses on both grid- (voltage stability) and building-level (thermal storage, PV curtailment) objectives. Constraints (thermal comfort, voltage, power factor) are enforced through external checks. GridLearn uses PettingZoo for multi-agent RL and leverages PandaPower for network simulation.

Gym-ANM (Henry and Ernst, 2021) offers single-agent active network management (ANM) for controlling distributed generation, storage, and voltage or line flows. It integrates with Stable Baselines3 for RL and provides penalty-based rewards for constraint handling (voltage, line capacity). Benchmark tasks compare with baselines (MPC, PPO, SAC) on scenarios such as wind-dominant nights, EV-heavy demand, and high-renewable scenarios.

PowerGym (Fan *et al.*, 2022) specifically targets VVC in distribution feeders. It includes standardized IEEE test systems (13-, 34-, 123-, and 8500-node), uses OpenDSS for multi-phase load flow, and enforces constraints (voltage bounds, SOC limits) via action projections. Different RL algorithms (PPO, SAC) are evaluated and compared for various control horizons and battery configurations.

**Multi-Application Platforms** DSS-SimPy-RL (Sahu *et al.*, 2023) combines power distribution tasks (network reconfiguration, VVC, CLR) with cyber network routing and congestion management. By interfacing OpenDSS (for power flow) and SimPy (for discrete-event cyber simulations), it creates a lightweight alternative to full co-simulation frameworks such as HELICS. The environment models both physical (voltage, capacity) and cyber (router queue limits, channel capacity) constraints as part of the reward and state monitoring, using metrics such as betweenness centrality to measure resilience. It integrates with standard RL libraries (DQN, PPO, A2C) and has been validated on IEEE distribution systems (13-bus, 34-bus, 123-bus, and 8500-node), with comparisons to random actions and spanning-tree heuristics.

SustainGym also integrates with established tools (ACN-Sim, EnergyPlus) and RL libraries (RLLib, Stable Baselines3), and it provides benchmarking against non-RL baselines, MPC, and standard RL algorithms (SAC, PPO). Because it offers dedicated modules for EV charging or building control, it can be briefly referenced in those respective sections, but its wide range of tasks and emphasis on $CO_2$-based objectives place it more naturally under "Multi-Domain Platforms."

SustainGym (Yeh *et al.*, 2024) unites several power- and sustainability-related tasks under a single environment, including EV charging, market bidding, data center scheduling, cogeneration dispatch, and building control, with a unique focus on realistic distribution shifts. Three of its five environments explicitly support multi-agent control, making it useful for cooperative scenarios such as building coordination or distributed EV charging. Constraints are enforced through penalty-based rewards rather than hard bounds. SustainGym integrates with established tools (ACN-Sim, EnergyPlus) and RL libraries (RLLib, Stable Baselines3), and it provides benchmarking against non-RL baselines, MPC, and standard RL algorithms (SAC, PPO).

**RL Integration and Benchmark Support** Most simulators adopt Gym-compatible interfaces, reducing the overhead for researchers using libraries like Stable Baselines and RLlib (Lee *et al.*, 2021; Pigott *et al.*, 2022; Vazquez-Canteli *et al.*, 2020; Biagioni *et al.*, 2022; Fan *et al.*, 2022; Yeh *et al.*, 2024; Henry and Ernst, 2021; Orfanoudakis *et al.*, 2024; Sahu *et al.*, 2023). Single-agent tasks commonly use standard Gym wrappers, while multi-agent scenarios rely on PettingZoo or RLlib's multi-agent extensions. The presence of built-in baselines (e.g., RBC or MPC) and standard test feeders further encourages reproducibility and consistent benchmarking across different RL algorithms. This integration reduces barriers to entry for newcomers, facilitating robust benchmarking against established algorithms.

**Scalability and Real-Time Performance** Many platforms highlight potential for large-scale or real-time studies, yet few provide thorough benchmarks of computational speed at increasing scales. Works such as (Vazquez-Canteli *et al.*, 2020; Pigott *et al.*, 2022; Biagioni *et al.*, 2022), and (Yeh *et al.*, 2024) mention potential efficiency optimizations (e.g., using pre-simulated data or distributed computing). (Henry and Ernst, 2021) offers timing comparisons for RL vs. MPC on moderate-scale tasks, and (Orfanoudakis *et al.*, 2024) highlights runtime evaluations up to 10,000 charging stations—illustrating linear growth at smaller scales and exponential growth beyond certain thresholds. Such granular performance reporting is vital for researchers aiming to adapt these simulators to real-time or large-scale deployments.

**Proprietary Simulation Environments in RL** Proprietary simulators such as DIgSILENT PowerFactory, PSS®E, and PSCAD offer high-fidelity modeling capturing detailed system dynamics, protection schemes, and device-specific behaviors. Researchers integrate these tools through Python APIs, COM interfaces, or co-simulation frameworks. PowerFactory connects to Python-based RL agents via socket communication or direct API calls, while PSS®E's Python interface (PSSPY) and PSCAD's automation scripts implement environments where agents receive measurements and return control actions.

Typical applications include voltage control using PowerFactory models for detailed network responses, frequency regulation using PSS®E for transmission grid dynamics, and transient/EMT studies using PSCAD for microgrid control and fault recovery. While providing unmatched accuracy and industry relevance, these tools present challenges including slower simulation speeds hindering rapid training, licensing requirements, and complex integration compared to open-source alternatives. Table 7.2 compares these approaches—proprietary simulators excel when research requires high accuracy and detailed modeling, while open-source environments facilitate rapid experimentation with lower computational overhead.

**Integration with Non-RL Methods** Although several platforms focus heavily on RL, some do facilitate comparisons with traditional controllers. For instance, Vazquez-Canteli *et al.* (2020) offer a rule-based baseline, while Lee *et al.* (2021) and Henry and Ernst (2021) integrate with mathematical programming tools (CVXPY) for MPC-based benchmarks. Orfanoudakis *et al.* (2024) include both heuristic baselines and commercial solver integration (Gurobi) for systematic performance comparisons. Going forward, adding more robust ties to mathematical programming frameworks and classical control methods

**Table 7.2:** Comparative Overview of Proprietary and Open-Source RL Environments for Power Systems

| Aspect | Proprietary Simulators | Open-Source Environments |
|---|---|---|
| **Examples** | DIgSILENT PowerFactory, PSS®E, PSCAD | See Table 7.1 |
| **Integration** | Python APIs, COM interfaces, co-simulation frameworks | Native Gym API wrappers; pure Python implementations |
| **Fidelity** | High-fidelity models capturing detailed device dynamics and protection schemes | Simplified or aggregated models; focus on computational efficiency |
| **Advantages** | Industry-grade accuracy; validated against real-world data | Fast simulation speeds; ease of prototyping and reproducibility |
| **Challenges** | Slower simulation; licensing costs; complex integration | Limited detail; may overlook low-level grid dynamics |

would strengthen each platform's utility as a comprehensive benchmarking tool. Furthermore, there is a strong need and potential to leverage that support to develop and test hybrid methods that combine RL with traditional methods such as MPC, which has demonstrated a strong potential in some of the existing benchmarks Khattar and Jin, 2023.

**Safety Handling and Operational Constraints** Constraint handling commonly depends on the following two strategies:

- Simulator Enforced: Action clipping or environment overrides when unsafe actions occur (e.g., CityLearn, GridLearn).
- Penalty-Based: Violations incur negative rewards, with no hard action blocking (e.g., PowerGridWorld, EV2Gym, Gym-ANM, PowerGym, SustainGym).

These approaches can be adapted for safe RL by exposing separate constraint signals in the environment's observation or info dictionary, allowing safe RL algorithms to track violations explicitly. However, none of the surveyed environments inherently implement a safe RL interface. Researchers still must modify reward structures or environment logic to ensure constraint feasibility throughout training.

# 8

# Open Challenges, Lessons Learned, and Future Directions

## Safety Guarantees During Learning and Operation

**Open Challenge:** *How can we ensure robust safety guarantees during both the learning process and real-time operation in power systems, without incurring prohibitive computation while adapting to changes?*

A "safety in depth" approach offers a promising direction: multiple protective layers can be combined or selected based on timescale and performance constraints. One layer focuses on provably safe policies: for instance, Jin and Lavaei (2020) and Gu *et al.* (2022) use stability certificates via IQCs to obtain a convex set of safe policy parameters, while Cui *et al.* (2023) enforce monotonicity of the policy. Another layer introduces real-time safety filters: Zhang *et al.* (2023c) employ DNN-assisted projection for voltage control, and Shi *et al.* (2023) use first-order approximations to reduce computation. A final layer adapts safety filters to shifting conditions, as Zhao *et al.* (2023) dynamically tune parameters to handle uncertainties, Wan *et al.* (2023) adjust CBF parameters for frequency control, and Zhang *et al.* (2023b) implement time-varying Lyapunov constraints in EV charging. These layers can also interact: for example, Chen *et al.* (2022) and Sun *et al.* (2024) show that logging both pre- and post-filtered actions accelerates learning of safer policies. Each layer must balance strict constraints (e.g., barrier functions) that limit exploration but guarantee near-zero violations, and softer, penalty-based methods that allow controlled violations but risk higher uncertainty. Moreover, combining model-free SRL (focused on policy parameters) with model-based methods (for runtime assurance) can naturally yield a hybrid system that maintains flexibility while strengthening safety.

> **Takeaway:** An integrated, multi-layered design can address varied safety requirements by combining innately safe policies, runtime corrections via optimization-based filters, and adaptive modules for evolving conditions. Yet, significant research gaps remain in unifying these tools to achieve both rigorous safety and the adaptability demanded by large-scale, dynamic power systems.

### Addressing Large-Scale Complexity

**Open Challenge:** *How can RL solutions manage the combinatorial explosion of nodes, actions, and data in large-scale power systems under real-world constraints such as partial observability, communication delays, and dynamic network topologies?*

Decomposition helps isolate complexity into smaller segments. Control-scope partitioning, such as bi-level design for local load agents and broader distribution operations (Zhang *et al.*, 2024a), or physics-based methods that exploit generator coherency (Kwon *et al.*, 2024), allow targeted control of inter-area oscillations without losing global coordination.

Managing sparse rewards and long horizons poses additional challenges, as seen in CLR (Chapter 6.5) or day-ahead scheduling. Shaping intermediate rewards (e.g., partial load restoration) or adopting hierarchical RL with sub-goals can accelerate training and boost final policy quality.

Topology-awareness improves scalability by embedding network structure in the control design. Graph-based methods (e.g., (Mu *et al.*, 2023; Jacob *et al.*, 2024)) preserve local relationships and cut down on parameter growth for tasks such as reconfiguration or VVC.

Distributed and multi-agent approaches (Chapter 4.2) further mitigate complexity by assigning local controllers to network subsets. However, they introduce challenges such as partial observability, nonstationarity, and coordination overhead. Efficient communication protocols (e.g., neighbor-to-neighbor with spatial discounting) and consensus mechanisms can curb excessive messaging, though delays, topology changes, and agent failures must be addressed (Liu and Wu, 2021; Guo *et al.*, 2023; Fan *et al.*, 2023b). Parallelization additionally streamlines training: collecting experience from multiple environments and sharing parameters reduces wall-clock time (Zhang *et al.*, 2023a; Shi *et al.*, 2023) and aligns well with the CTDE framework (Chapter 4).

> **Takeaway:** Segmenting large power grids by control scope, physics-based structure, or timescales eases complexity, while reward-shaping, topology-aware design, coordination protocols, and parallelization keep training and execution scalable. However, real-world communication delays, agent failures, and non-stationary conditions pose challenges for reliable operation at scale.

### Rethinking Safety in Uncertain and Nonstationary Environments

**Open Challenge:** *How can RL methods achieve "open-world" safety in power systems, where frequent distributional shifts and extreme black-swan events demand the agent to remain effective and adaptive throughout unpredictable, off-nominal operating states.*

Modern safe RL approaches often view safety as a fixed constraint: the agent simply avoids known violations in a stationary, "closed-world" environment. Yet real-world power grids are anything but static, with demand shifts, renewable volatility (especially in low-inertia systems), environmental disruptions, and potential cyberattacks. True safety necessitates robust or risk-aware *generalization* that extends beyond finite training scenarios and prepares the agent for unseen "black-swan" events.

A vital shift is to regard *safety as a performance metric* (Figure 8.1). Rather than aiming to be "safe enough" in a narrow set of conditions, an RL policy actively seeks to excel in safe behavior *across* distributional

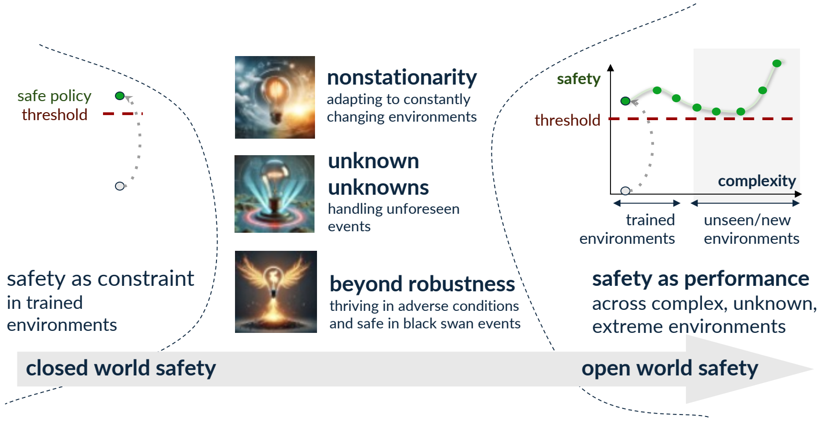


**Figure 8.1:** A conceptual shift from closed-world safety—treating safety as a constraint in familiar, trained conditions—toward open-world safety, where the system must adapt to nonstationarity, unforeseen events, and extreme scenarios by treating safety itself as a performance objective.

shifts. Achieving this shift requires addressing nonstationarity, accounting for unforeseen or partially observed threats. Achieving this pivot involves coping with nonstationarity, anticipating partially observed threats, and refining safe control strategies in real time *before* failures escalate.

Meta-safe RL offers a promising framework for tackling these challenges, with a meta-learner continuously improving how a base learner interacts with new tasks or environmental states (c.f., Fig. 2.2 for an illustration). By refining strategies on the fly, the system acquires *adaptive safety*, swiftly assimilating new constraints or hazard information while preserving functional performance.

This approach aligns with the idea of *antifragility* (Taleb, 2014), which pushes beyond mere robustness by treating severe disruptions and rare events as opportunities to improve rather than as temporary setbacks. As discussed in (Jin, 2024), computational antifragility can be pursued through meta-learning that accelerates safe adaptation with each disturbance, POMDPs and continual learning that incorporate partial observability, multi-objective RL that balances safety with other grid goals, and foundation models and pretraining that enable in-context learning. By integrating these paradigms, an RL agent can adapt *within* an extreme event, avoid catastrophic failures before they unfold, and emerge more capable of tackling future contingencies.

> **Takeaway:** Safe RL must move from static, constraint-focused methods to approaches that view safety as a core performance goal under distributional shifts. In doing so, agents become able to adapt and even improve their safe behavior under disruptive conditions, aligning with antifragile principles that favor continual learning and mid-event refinement to avert catastrophic failures.

### Leveraging Foundation Models for Enhanced Safety and Scalability

**Open Challenge:** *How can large-scale, pretrained foundation models be effectively integrated into power system control to enhance safety, scalability, and adaptability—across simulation environments, model-based planning, and real-time operation—without compromising strict operational constraints?*

Foundation models are large-scale neural networks trained on diverse, extensive datasets that can be adapted to a wide range of downstream tasks with minimal fine-tuning. While these models have transformed fields such as natural language processing and computer vision, their application to power systems is still emerging. These models show promise for learning transferable representations that could address complex power grid challenges from ontingency analysis and forecasting to control operations and cybersecurity (Bommasani *et al.*, 2021; Hamann *et al.*, 2024). As demonstrated in DeepSeek-r1 (Guo *et al.*, 2025), foundation models can be enhanced through RL with sparse, rule-based rewards—a crucial advantage for power systems where data is scarce but constraints are well-defined.

Recent applications include large language models (LLMs) for optimization auto-formalism that translate natural language into solvable optimization formats (Jin *et al.*, 2024), and PowerPM (Tu *et al.*, 2024), a pioneering foundation model for electricity time series that captures both temporal dependencies and hierarchical relationships, demonstrating strong performance across 44 downstream power system tasks. In simulation environments, these models could enable fast digital twins for accelerated control policy training.

Despite promising advances, significant challenges persist: bridging the sim-to-real gap, ensuring real-time inference under strict latency requirements, and achieving formal safety certification. Embedding physics-informed constraints and safe RL frameworks with runtime safety layers would help maintain critical system limits during online adaptation (Chapter 5). Integration with legacy infrastructure requires careful design to preserve operator trust and regulatory compliance.

> **Takeaway:** Foundation models offer transformative potential for scalable, adaptable power system control, but critical challenges in real-time performance, safety guarantees, and seamless integration with existing systems must be addressed before full operational deployment in live grid environments.

### Integration with Existing Infrastructure and Operational AI

**Open Challenge:** *How can advanced safe RL methods be effectively integrated into legacy power system infrastructures—spanning local droop controllers, discrete tap changers, SCADA networks, and regulatory frameworks—without disrupting trusted operational routines or requiring prohibitively extensive overhauls?*

A central theme across many safe RL deployments is the decision to *retain* proven baseline controllers and architectures, with RL providing higher-level coordination or advisory signals. In frequency regulation, for instance, Cui *et al.* (2023) and Kwon *et al.* (2023) adjust the setpoints of existing droop loops rather than replacing them outright, thus preserving fundamental control structures while introducing an adaptive learning layer on top. This same pattern appears in volt-VAR control, where Sun *et al.* (2024) position RL as a decision-support module rather than an autonomous controller, thereby balancing system-wide optimization with local responsiveness.

Communication demands vary considerably depending on the scope and timescale of control. In frequency regulation, some works require minimal local measurements (e.g., (Shuai *et al.*, 2024; Cui *et al.*, 2023)) or short-range communications, whereas others employ more comprehensive real-time data exchanges of tie-line flows and generator statuses (Kwon *et al.*, 2023). Update frequencies range from sub-second intervals (100Hz) for fine-grained droop setpoint adjustments to multi-minute intervals (Xia *et al.*, 2022). Approaches in volt-VAR control use time-synchronization or neighbor-based averaging (Zhang *et al.*, 2023b; Guo *et al.*, 2023), allowing effective coordination even with limited data links.

On the hardware and computational side, many RL-based frequency control strategies (e.g., (Xia *et al.*, 2022; Shuai *et al.*, 2024)) run on standard computing platforms, yet demands vary substantially. Simpler control laws compute new setpoints within milliseconds, while advanced approaches employing large neural networks or quadratic programming may require more time or parallel processing. For critical load restoration, RL typically connects directly to dispatchable generation and storage via existing DER monitoring systems, echoing outage management procedures that handle fault messages and system topologies.

Industry trust and formal certification remain pressing concerns. Demonstrating compliance with grid codes (NERC BAL-003-1, ENTSO-E guidelines) and meeting strict frequency-recovery timelines encourages a hybrid approach in which RL solutions coexist with fallback controllers, robust safety verification, and hardware-in-the-loop tests. For EV aggregation, Zhang *et al.* (2023a) align RL strategies with FERC Order No. 2222, detailing incremental infrastructure modifications such as continuous charging-rate controls. In all these contexts, the overarching lesson is that *modern RL must enhance or complement existing control methods rather than replace them*, thereby reducing adoption barriers while preserving reliability.

A promising future direction is the operational deployment of RL-based control systems in live power grids. While many RL approaches have been validated in simulation or pilot studies, transitioning them into production environments necessitates stringent safety verification, real-time adaptability, and seamless integration with legacy systems. Early deployments—such as the High Performance Adaptive Deep-Reinforcement-Learning-based Real-time Emergency Control (HADREC) emergency controller (Chen, 2023) and Dubai Electricity and Water Authority (DEWA)'s RL-based gas turbine auto-tuning—

demonstrate the potential of operational AI. Additionally, pilots in wind farm control and demand-side management illustrate that, with appropriate hybrid architectures (combining offline training on digital twins with adaptive, real-time safety filters), safe and multi-agent RL can effectively enhance grid performance without compromising system reliability.

> **Key Takeaways.** Safe RL thrives best when it integrates seamlessly with established operating routines, from droop or PI regulators in frequency control to tap-changer logics in volt-VAR tasks. Rather than discarding proven solutions, most implementations overlay RL-driven coordination or advisory signals on top of legacy systems, ensure robust data pathways, and satisfy industry certification demands through fallback mechanisms, real-time constraints, and conservative incremental deployment.

# Acknowledgements

This work was supported by the National Science Foundation under Grant No. 2331775. The author would like to thank Dr. Javad Lavaei, Dr. Chen-Ching Liu, Dr. Costas Spanos, and Dr. Alberto Sangiovanni-Vincentelli for their invaluable guidance and support. The authors would also like to thank Dr. Nanpeng Yu, Dr. Qianwen Xu, Dr. Ali Mehrizi-Sani, Dr. Lamine Mili, and the anonymous reviewer for their review and constructive feedback. The authors gratefully acknowledge the contributions of the members of the ROLE research group and the feedback from graduate students in Power and Energy Center @ VT.

# References

Abdeen, Z. ul, X. Zhang, W. Gill, and M. Jin. (2024). “Enhancing Distribution System Resilience: A First-Order Meta-RL algorithm for Critical Load Restoration”. In: *2024 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE. 129–134.

Achiam, J., D. Held, A. Tamar, and P. Abbeel. (2017). “Constrained policy optimization”. In: *International conference on machine learning*. PMLR. 22–31.

Ackermann, J., V. Gabler, T. Osa, and M. Sugiyama. (2019). “Reducing overestimation bias in multi-agent domains using double centralized critics”. *arXiv preprint arXiv:1910.01465*.

Altman, E. (2021). *Constrained Markov decision processes*. Routledge.

Arcak, M., C. Meissen, and A. Packard. (2018). “Networks of Dissipative Systems”.

Bai, N., Q. Wang, Z. Duan, and G. Chen. (2024). “Distributed Optimal Consensus Control of Constrained Multi-Agent Systems: A Non-Separable Optimization Perspective”. *IEEE Transactions on Automatic Control*.

Bansal, S., M. Chen, S. Herbert, and C. J. Tomlin. (2017). “Hamilton-jacobi reachability: A brief overview and recent advances”. In: *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE. 2242–2253.

Beikbabaei, M., B. M. Kwiatkowski, and A. Mehrizi-Sani. (2025). "Model-Free Resilient Grid-Forming and Grid-Following Inverter Control Against Cyberattacks Using Reinforcement Learning". *Electronics.* 14(2): 288.

Biagioni, D., X. Zhang, D. Wald, D. Vaidhynathan, R. Chintala, J. King, and A. S. Zamzam. (2022). "Powergridworld: A framework for multi-agent reinforcement learning in power systems". In: *Proceedings of the thirteenth ACM international conference on future energy systems.* 565–570.

Biswas, A., M. Acquarone, H. Wang, F. Miretti, D. A. Misul, and A. Emadi. (2024). "Safe Reinforcement Learning for Energy Management of Electrified Vehicle with Novel Physics-Informed Exploration Strategy". *IEEE Transactions on Transportation Electrification.*

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.* (2021). "On the opportunities and risks of foundation models". *arXiv preprint arXiv:2108.07258.*

Brunke, L., M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig. (2022). "Safe learning in robotics: From learning-based control to safe reinforcement learning". *Annual Review of Control, Robotics, and Autonomous Systems.* 5(1): 411–444.

Chang, Y.-C., N. Roohi, and S. Gao. (2019). "Neural lyapunov control". *Advances in neural information processing systems.* 32.

Chen, D., K. Chen, Z. Li, T. Chu, R. Yao, F. Qiu, and K. Lin. (2021). "Powernet: Multi-agent deep reinforcement learning for scalable powergrid control". *IEEE Transactions on Power Systems.* 37(2): 1007–1017.

Chen, P., S. Liu, X. Wang, and I. Kamwa. (2022). "Physics-shielded multi-agent deep reinforcement learning for safe active voltage control with photovoltaic/battery energy storage systems". *IEEE Transactions on Smart Grid.* 14(4): 2656–2667.

Chen, Y., S. Huang, F. Liu, Z. Wang, and X. Sun. (2018). "Evaluation of reinforcement learning-based false data injection attack to automatic voltage control". *IEEE Transactions on Smart Grid.* 10(2): 2158–2169.

Chen, Y., Y. Liu, H. Yin, Z. Tang, G. Qiu, and J. Liu. (2024). “Multiagent Soft Actor–Critic Learning for Distributed ESS Enabled Robust Voltage Regulation of Active Distribution Grids”. *IEEE Transactions on Industrial Informatics.*

Chen, Y. (2023). “High Performance Adaptive Deep-Reinforcement-Learning-based Real-time Emergency Control (HADREC) to Enhance Power Grid Resilience in Stochastic Environment Final Report”. *Tech. rep.* Pacific Northwest National Laboratory (PNNL), Richland, WA (United States).

Cui, W., Y. Jiang, and B. Zhang. (2023). “Reinforcement learning for optimal primary frequency control: A Lyapunov approach”. *IEEE Transactions on Power Systems.* 38(2): 1676–1688.

Du, Y. and D. Wu. (2022). “Deep reinforcement learning from demonstrations to assist service restoration in islanded microgrids”. *IEEE Transactions on Sustainable Energy.* 13(2): 1062–1072.

Fan, B., X. Liu, G. Xiao, Y. Kang, D. Wang, and P. Wang. (2023a). “Attention-Based Multi-Agent Graph Reinforcement Learning for Service Restoration”. *IEEE Transactions on Artificial Intelligence.*

Fan, B., X. Liu, G. Xiao, Y. Xu, X. Yang, and P. Wang. (2024). “A Memory-Based Graph Reinforcement Learning Method for Critical Load Restoration With Uncertainties of Distributed Energy Resource”. *IEEE Transactions on Smart Grid.*

Fan, T.-H., X. Y. Lee, and Y. Wang. (2022). “Powergym: A reinforcement learning environment for volt-var control in power distribution systems”. In: *Learning for Dynamics and Control Conference.* PMLR. 21–33.

Fan, Z., W. Zhang, and W. Liu. (2023b). “Multi-agent deep reinforcement learning based distributed optimal generation control of DC microgrids”. *IEEE Transactions on Smart Grid.*

Feng, J., Y. Shi, G. Qu, S. H. Low, A. Anandkumar, and A. Wierman. (2023). “Stability constrained reinforcement learning for decentralized real-time voltage control”. *IEEE Transactions on Control of Network Systems.*

Foerster, J., I. A. Assael, N. De Freitas, and S. Whiteson. (2016). “Learning to communicate with deep multi-agent reinforcement learning”. *Advances in neural information processing systems.* 29.

Foerster, J., G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson. (2018). "Counterfactual multi-agent policy gradients". In: *Proceedings of the AAAI conference on artificial intelligence.* Vol. 32. No. 1.

Gao, Y., W. Wang, and N. Yu. (2021). "Consensus multi-agent reinforcement learning for volt-var control in power distribution networks". *IEEE Transactions on Smart Grid.* 12(4): 3594–3604.

Gao, Y. and N. Yu. (2022). "Model-augmented safe reinforcement learning for Volt-VAR control in power distribution networks". *Applied Energy.* 313: 118762.

Gautam, M. (2023). "Deep Reinforcement learning for resilient power and energy systems: Progress, prospects, and future avenues". *Electricity.* 4(4): 336–380.

Gu, F., H. Yin, L. El Ghaoui, M. Arcak, P. Seiler, and M. Jin. (2022). "Recurrent neural network controllers synthesis with stability guarantees for partially observed systems". In: *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 36. No. 5. 5385–5394.

Gu, S., J. G. Kuba, Y. Chen, Y. Du, L. Yang, A. Knoll, and Y. Yang. (2023). "Safe multi-agent reinforcement learning for multi-robot control". *Artificial Intelligence.* 319: 103905.

Gu, S., B. Sel, Y. Ding, L. Wang, Q. Lin, M. Jin, and A. Knoll. (2024a). "Balance reward and safety optimization for safe reinforcement learning: A perspective of gradient manipulation". In: *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 38. No. 19. 21099–21106.

Gu, S., L. Shi, Y. Ding, A. Knoll, C. Spanos, A. Wierman, and M. Jin. (2024b). "Enhancing Efficiency of Safe Reinforcement Learning via Sample Manipulation". *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS).*

Guo, D., D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, *et al.* (2025). "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning". *arXiv preprint arXiv:2501.12948.*

Guo, G., M. Zhang, Y. Gong, and Q. Xu. (2023). "Safe multi-agent deep reinforcement learning for real-time decentralized control of inverter based renewable energy resources considering communication delay". *Applied Energy.* 349: 121648.

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

Habib, B., E. Isufi, W. van Breda, A. Jongepier, and J. L. Cremer. (2023). “Deep statistical solver for distribution system state estimation”. *IEEE Transactions on Power Systems*. 39(2): 4039–4050.

Hamann, H. F., B. Gjorgiev, T. Brunschwiler, L. S. Martins, A. Puech, A. Varbella, J. Weiss, J. Bernabe-Moreno, A. B. Massé, S. L. Choi, *et al.* (2024). “Foundation models for the electric power grid”. *Joule*. 8(12): 3245–3258.

He, H., J. Boyd-Graber, K. Kwok, and H. Daumé III. (2016). “Opponent modeling in deep reinforcement learning”. In: *International conference on machine learning*. PMLR. 1804–1813.

Henry, R. and D. Ernst. (2021). “Gym-ANM: Reinforcement learning environments for active network management tasks in electricity distribution systems”. *Energy and AI*. 5: 100092.

Hernandez-Leal, P., M. Kaisers, T. Baarslag, and E. M. De Cote. (2017). “A survey of learning in multiagent environments: Dealing with non-stationarity”. *arXiv preprint arXiv:1707.09183*.

Hewing, L., K. P. Wabersich, M. Menner, and M. N. Zeilinger. (2020). “Learning-based model predictive control: Toward safe learning in control”. *Annual Review of Control, Robotics, and Autonomous Systems*. 3(1): 269–296.

Hobbs, K. L., M. L. Mote, M. C. Abate, S. D. Coogan, and E. M. Feron. (2023). “Runtime assurance for safety-critical systems: An introduction to safety filtering approaches for complex control systems”. *IEEE Control Systems Magazine*. 43(2): 28–65.

Hong, L., M. Wu, Y. Wang, M. Shahidehpour, Z. Chen, and Z. Yan. (2024). “MADRL-Based DSO-Customer Coordinated Bi-Level Volt/-VAR Optimization Method for Power Distribution Networks”. *IEEE Transactions on Sustainable Energy*.

Hong, S.-H. and H.-S. Lee. (2023). “Robust energy management system with safe reinforcement learning using short-horizon forecasts”. *IEEE Transactions on Smart Grid*. 14(3): 2485–2488.

Hou, S., E. M. Salazar, P. Palensky, Q. Chen, and P. P. Vergara. (2024). "A mix-integer programming based deep reinforcement learning framework for optimal dispatch of energy storage system in distribution networks". *Journal of Modern Power Systems and Clean Energy*.

Hsu, K.-C., H. Hu, and J. F. Fisac. (2023). "The safety filter: A unified view of safety-critical control in autonomous systems". *Annual Review of Control, Robotics, and Autonomous Systems*. 7.

Hu, C., J. Yan, and X. Liu. (2022a). "Reinforcement learning-based adaptive feature boosting for smart grid intrusion detection". *IEEE Transactions on Smart Grid*. 14(4): 3150–3163.

Hu, D., Z. Ye, Y. Gao, Z. Ye, Y. Peng, and N. Yu. (2022b). "Multi-agent deep reinforcement learning for voltage control with coordinated active and reactive power optimization". *IEEE Transactions on Smart Grid*. 13(6): 4873–4886.

Hu, J., Y. Ye, Y. Wu, P. Zhao, and L. Liu. (2024). "Rethinking Safe Policy Learning for Complex Constraints Satisfaction: A Glimpse in Real-Time Security Constrained Economic Dispatch Integrating Energy Storage Units". *IEEE Transactions on Power Systems*.

Hu, L., C. Wu, and W. Pan. (2020). "Lyapunov-based reinforcement learning state estimator". *arXiv preprint arXiv:2010.13529*.

Iqbal, S. and F. Sha. (2019). "Actor-attention-critic for multi-agent reinforcement learning". In: *International conference on machine learning*. PMLR. 2961–2970.

Jacob, R. A., S. Paul, S. Chowdhury, Y. R. Gel, and J. Zhang. (2024). "Real-time outage management in active distribution networks using reinforcement learning over graphs". *Nature Communications*. 15(1): 4766.

Jeon, S., H. T. Nguyen, and D.-H. Choi. (2023). "Safety-Integrated Online Deep Reinforcement Learning for Mobile Energy Storage System Scheduling and Volt/VAR Control in Power Distribution Networks". *IEEE Access*. 11: 34440–34455.

Jiang, J. and Z. Lu. (2018). "Learning attentional communication for multi-agent cooperation". *Advances in neural information processing systems*. 31.

Jiang, Y., Q. Ye, B. Sun, Y. Wu, and D. H. Tsang. (2021). “Data-driven coordinated charging for electric vehicles with continuous charging rates: A deep policy gradient approach”. *IEEE Internet of Things Journal.* 9(14): 12395–12412.

Jin, M. (2024). “Preparing for Black Swans: The Antifragility Imperative for Machine Learning”. *arXiv preprint arXiv:2405.11397.*

Jin, M. and J. Lavaei. (2020). “Stability-certified reinforcement learning: A control-theoretic perspective”. *IEEE Access.* 8: 229086–229100.

Jin, M., J. Lavaei, S. Sojoudi, and R. Baldick. (2020). “Boundary defense against cyber threat for power system state estimation”. *IEEE Transactions on Information Forensics and Security.* 16: 1752–1767.

Jin, M., B. Sel, F. Hardeep, and W. Yin. (2024). “Democratizing energy management with llm-assisted optimization autoformalism”. In: *2024 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm).* IEEE. 258–263.

Kabir, F., N. Yu, Y. Gao, and W. Wang. (2023). “Deep reinforcement learning-based two-timescale Volt-VAR control with degradation-aware smart inverters in power distribution systems”. *Applied Energy.* 335: 120629.

Kakade, S. and J. Langford. (2002). “Approximately optimal approximate reinforcement learning”. In: *Proceedings of the Nineteenth International Conference on Machine Learning.* 267–274.

Kesici, M., B. Pal, and G. Yang. (2024). “Detection of false data injection attacks in distribution networks: A vertical federated learning approach”. *IEEE Transactions on Smart Grid.*

Khattar, V., Y. Ding, B. Sel, J. Lavaei, and M. Jin. (2023). “A CMDP-within-online framework for Meta-Safe Reinforcement Learning”. In: *The Eleventh International Conference on Learning Representations.*

Khattar, V. and M. Jin. (2023). “Winning the CityLearn challenge: adaptive optimization with evolutionary search under trajectory-based guidance”. In: *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 37. No. 12. 14286–14294.

Khattar, V., Y. Yao, F. Ding, and M. Jin. (2025). “Distribution Grid Critical Load Restoration under Uncertain Topology Changes via a Hierarchical Multi-Agent Reinforcement Learning Approach”. In: *IEEE Power & Energy Society General Meeting.* IEEE. 1–5.

Kurt, M. N., O. Ogundijo, C. Li, and X. Wang. (2018). “Online cyber-attack detection in smart grid: A reinforcement learning approach”. *IEEE Transactions on Smart Grid.* 10(5): 5174–5185.

Kwon, K.-B., R. R. Hossain, S. Mukherjee, K. Chatterjee, S. Kundu, S. Nekkalapu, and M. Elizondo. (2024). “Coherency-Aware Learning Control of Inverter-Dominated Grids: A Distributed Risk-Constrained Approach”. *IEEE Control Systems Letters.*

Kwon, K.-b., S. Mukherjee, T. L. Vu, and H. Zhu. (2023). “Risk-Constrained Reinforcement Learning for Inverter-Dominated Power System Controls”. *IEEE Control Systems Letters.*

Lee, Z. J., S. Sharma, D. Johansson, and S. H. Low. (2021). “ACN-Sim: An Open-Source Simulator for Data-Driven Electric Vehicle Charging Research”. *IEEE Transactions on Smart Grid.* 12(6): 5113–5123.

Li, C. and Z. Wu. (2024). “F-DQN: an optimized DQN for decision-making of generator start-up sequence after blackout”. *Applied Intelligence*: 1–15.

Li, H. and H. He. (2022). “Learning to operate distribution networks with safe deep reinforcement learning”. *IEEE Transactions on Smart Grid.* 13(3): 1860–1872.

Li, H., Z. Wan, and H. He. (2019). “Constrained EV charging scheduling based on safe deep reinforcement learning”. *IEEE Transactions on Smart Grid.* 11(3): 2427–2439.

Li, S., W. Hu, D. Cao, Z. Chen, Q. Huang, F. Blaabjerg, and K. Liao. (2023). “Physics-model-free heat-electricity energy management of multiple microgrids based on surrogate model-enabled multi-agent deep reinforcement learning”. *Applied Energy.* 346: 121359.

Li, S., W. Hu, D. Cao, J. Hu, Q. Huang, Z. Chen, and F. Blaabjerg. (2024). “A Novel MADRL with Spatial-Temporal Pattern Capturing Ability for Robust Decentralized Control of Multiple Microgrids under Anomalous Measurements”. *IEEE Transactions on Sustainable Energy.*

Lin, Y., W. Li, H. Zha, and B. Wang. (2024). “Information Design in Multi-Agent Reinforcement Learning”. *Advances in Neural Information Processing Systems*. 36.

Liu, H. and W. Wu. (2021). “Online multi-agent reinforcement learning for decentralized inverter-based volt-var control”. *IEEE Transactions on Smart Grid*. 12(4): 2980–2990.

Liu, L., N. Shi, D. Wang, Z. Ma, Z. Wang, M. J. Reno, and J. A. Azzolini. (2024a). “Voltage calculations in secondary distribution networks via physics-inspired neural network using smart meter data”. *IEEE Transactions on Smart Grid*.

Liu, X., Q. Jiao, S. Qiao, Z. Yan, S. Wen, and P. Wang. (2024b). “A Hybrid Monotonic Neural Network Approach for Multi-Area Power System Load Frequency Control Against FGSM Attack”. *IEEE Transactions on Circuits and Systems II: Express Briefs*.

Lohmiller, W. and J.-J. E. Slotine. (1998). “On contraction analysis for non-linear systems”. *Automatica*. 34(6): 683–696.

Lowe, R., Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch. (2017). “Multi-agent actor-critic for mixed cooperative-competitive environments”. *Advances in neural information processing systems*. 30.

Luo, S., M. Jin, R. Han, and K. Sun. (2022). “Anti-jamming based on Reinforcement Learning in Power System Sensing Network”. In: *Proceedings of the 2022 12th International Conference on Communication and Network Security*. 180–184.

Ma, D., Y. Wang, and S. Wu. (2024). “Against Jamming Attack in Wireless Communication Networks: A Reinforcement Learning Approach”. *Electronics*. 13(7): 1209.

Maiti, S. and S. Dey. (2024). “Smart Grid Security: A Verified Deep Reinforcement Learning Framework to Counter Cyber-Physical Attacks”. *arXiv preprint arXiv:2409.15757*.

Megretski, A. and A. Rantzer. (1997). “System analysis via integral quadratic constraints”. *IEEE transactions on automatic control*. 42(6): 819–830.

Moradi, M., S. Panahi, Z.-M. Zhai, Y. Weng, J. Dirkman, and Y.-C. Lai. (2024). “Heterogeneous reinforcement learning for defending power grids against attacks”. *APL Machine Learning*. 2(2).

Moradi, M., Y. Weng, J. Dirkman, and Y.-C. Lai. (2023). "Preferential cyber defense for power grids". *PRX Energy.* 2(4): 043007.

Mouyart, M., G. Medeiros Machado, and J.-Y. Jun. (2023). "A multi-agent intrusion detection system optimized by a deep reinforcement learning approach with a dataset enlarged using a generative model to reduce the bias effect". *Journal of Sensor and Actuator Networks.* 12(5): 68.

Mu, C., Z. Liu, J. Yan, H. Jia, and X. Zhang. (2023). "Graph multi-agent reinforcement learning for inverter-based active voltage control". *IEEE Transactions on Smart Grid.*

Mukherjee, S., R. R. Hossain, S. M. Mohiuddin, Y. Liu, W. Du, V. Adetola, R. A. Jinsiwale, Q. Huang, T. Yin, and A. Singhal. (2024). "Resilient Control of Networked Microgrids using Vertical Federated Reinforcement Learning: Designs and Real-Time Test-Bed Validations". *IEEE Transactions on Smart Grid.*

Nagy, Z., J. R. Vázquez-Canteli, S. Dey, and G. Henze. (2021). "The citylearn challenge 2021". In: *Proceedings of the 8th ACM international conference on systems for energy-efficient buildings, cities, and transportation.* 218–219.

Nguyen, H. T. and D.-H. Choi. (2022). "Three-stage inverter-based peak shaving and Volt-VAR control in active distribution networks using online safe deep reinforcement learning". *IEEE Transactions on Smart Grid.* 13(4): 3266–3277.

Oliehoek, F. A., C. Amato, *et al.* (2016). *A concise introduction to decentralized POMDPs.* Vol. 1. Springer.

Omidshafiei, S., J. Pazis, C. Amato, J. P. How, and J. Vian. (2017). "Deep decentralized multi-task multi-agent reinforcement learning under partial observability". In: *International Conference on Machine Learning.* PMLR. 2681–2690.

Orfanoudakis, S., C. Diaz-Londono, Y. E. Yılmaz, P. Palensky, and P. P. Vergara. (2024). "Ev2gym: A flexible v2g simulator for ev smart charging research and benchmarking". *IEEE Transactions on Intelligent Transportation Systems.*

Paesschesoone, S., N. Kayedpour, C. Manna, and G. Crevecoeur. (2024). "Reinforcement learning for an enhanced energy flexibility controller incorporating predictive safety filter and adaptive policy updates". *Applied Energy*. 368: 123507.

Pigott, A., C. Crozier, K. Baker, and Z. Nagy. (2022). "Gridlearn: Multiagent reinforcement learning for grid-aware building energy management". *Electric power systems research*. 213: 108521.

Qin, Z., K. Zhang, Y. Chen, J. Chen, and C. Fan. (2021). "Learning safe multi-agent control with decentralized neural barrier certificates". *arXiv preprint arXiv:2101.05436*.

Qiu, D., Z. Dong, X. Zhang, Y. Wang, and G. Strbac. (2022). "Safe reinforcement learning for real-time automatic control in a smart energy-hub". *Applied Energy*. 309: 118403.

Qiu, D., Y. Wang, J. Wang, N. Zhang, G. Strbac, and C. Kang. (2023). "Resilience-Oriented Coordination of Networked Microgrids: A Shapley Q-Value Learning Approach". *IEEE transactions on power systems*. 39(2): 3401–3416.

Rashid, T., M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson. (2020). "Monotonic value function factorisation for deep multi-agent reinforcement learning". *Journal of Machine Learning Research*. 21(178): 1–51.

Ray, A., J. Achiam, and D. Amodei. (2019). "Benchmarking safe exploration in deep reinforcement learning". *arXiv preprint arXiv:1910.01708*. 7(1): 2.

Sahu, A., V. Venkatraman, and R. Macwan. (2023). "Reinforcement learning environment for cyber-resilient power distribution system". *IEEE Access*. 11: 127216–127228.

Salamat, B., G. Elsbacher, A. M. Tonello, and L. Belzner. (2023). "Model-free distributed reinforcement learning state estimation of a dynamical system using integral value functions". *IEEE Open Journal of Control Systems*. 2: 70–78.

Schierman, J. D., M. D. DeVore, N. D. Richards, N. Gandhi, J. K. Cooper, K. R. Horneman, S. Stoller, and S. Smolka. (2015). "Runtime assurance framework development for highly adaptive flight control systems". *Barron Associates, Inc. Charlottesville, Tech. Rep.*

Seiler, P. (2014). “Stability analysis with dissipation inequalities and integral quadratic constraints”. *IEEE Transactions on Automatic Control*. 60(6): 1704–1709.

Sethi, K., Y. V. Madhav, R. Kumar, and P. Bera. (2021). “Attention based multi-agent intrusion detection systems using reinforcement learning”. *Journal of Information Security and Applications*. 61: 102923.

Seto, D., B. Krogh, L. Sha, and A. Chutinan. (1998). “The simplex architecture for safe online control system upgrades”. In: *Proceedings of the 1998 American Control Conference*. Vol. 6. IEEE. 3504–3508.

Shengren, H., P. P. Vergara, E. M. S. Duque, and P. Palensky. (2023). “Optimal energy system scheduling using a constraint-aware reinforcement learning algorithm”. *International Journal of Electrical Power & Energy Systems*. 152: 109230.

Shi, Y., M. Feng, M. Wang, W. Zhou, and H. Li. (2023). “Multi-Agent Reinforcement Learning with Safety Layer for Active Voltage Control”. In: *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 1533–1541.

Shuai, H., B. She, J. Wang, and F. Li. (2024). “Safe Reinforcement Learning for Grid-Forming Inverter Based Frequency Regulation with Stability Guarantee”. *Journal of Modern Power Systems and Clean Energy*.

Si, R., S. Chen, J. Zhang, J. Xu, and L. Zhang. (2024). “A multi-agent reinforcement learning method for distribution system restoration considering dynamic network reconfiguration”. *Applied Energy*. 372: 123625.

Sootla, A., A. I. Cowen-Rivers, T. Jafferjee, Z. Wang, D. H. Mguni, J. Wang, and H. Ammar. (2022). “Sauté rl: Almost surely safe reinforcement learning using state augmentation”. In: *International Conference on Machine Learning*. PMLR. 20423–20443.

Sukhbaatar, S., R. Fergus, *et al.* (2016). “Learning multiagent communication with backpropagation”. *Advances in neural information processing systems*. 29.

Sun, X., Z. Xu, J. Qiu, H. Liu, H. Wu, and Y. Tao. (2024). "Optimal volt/var control for unbalanced distribution networks with human-in-the-loop deep reinforcement learning". *IEEE Transactions on Smart Grid*.

Sun, Z. and T. Lu. (2024). "Collaborative operation optimization of distribution system and virtual power plants using multi-agent deep reinforcement learning with parameter-sharing mechanism". *IET Generation, Transmission & Distribution*. 18(1): 39–49.

Sunehag, P., G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, *et al.* (2018). "Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward". In: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 2085–2087.

Taleb, N. N. (2014). *Antifragile: Things that gain from disorder*. Vol. 3. Random House Trade Paperbacks.

Tan, X. and D. V. Dimarogonas. (2021). "Distributed implementation of control barrier functions for multi-agent systems". *IEEE Control Systems Letters*. 6: 1879–1884.

Tan, X., C. Liu, K. H. Johansson, and D. V. Dimarogonas. (2024). "A continuous-time violation-free multi-agent optimization algorithm and its applications to safe distributed control". *arXiv preprint arXiv:2404.07571*.

Tu, S., Y. Zhang, J. Zhang, Z. Fu, Y. Zhang, and Y. Yang. (2024). "Powerpm: Foundation model for power systems". *Advances in Neural Information Processing Systems*. 37: 115233–115260.

Vazquez-Canteli, J. R., S. Dey, G. Henze, and Z. Nagy. (2020). "CityLearn: Standardizing research in multi-agent reinforcement learning for demand response and urban energy management". *arXiv preprint arXiv:2012.10504*.

Vu, L., T. Vu, T. L. Vu, and A. Srivastava. (2023). "Multi-agent deep reinforcement learning for distributed load restoration". *IEEE Transactions on Smart Grid*.

Wan, X., M. Sun, B. Chen, Z. Chu, and F. Teng. (2023). "AdapSafe: adaptive and safe-certified deep reinforcement learning-based frequency control for carbon-neutral power systems". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. No. 4. 5294–5302.

Wang, J., Y. Zhang, T.-K. Kim, and Y. Gu. (2020). "Shapley Q-value: A local reward approach to solve global reward games". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 05. 7285–7292.

Wang, L., S. Zhang, Y. Zhou, C. Fan, P. Zhang, and Y. A. Shamash. (2023a). "Physics-informed, safety and stability certified neural control for uncertain networked microgrids". *IEEE Transactions on Smart Grid*.

Wang, W., N. Yu, Y. Gao, and J. Shi. (2019). "Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems". *IEEE Transactions on Smart Grid*. 11(4): 3008–3018.

Wang, Y., D. Qiu, M. Sun, G. Strbac, and Z. Gao. (2023b). "Secure energy management of multi-energy microgrid: A physical-informed safe reinforcement learning approach". *Applied Energy*. 335: 120759.

Wang, Y., D. Qiu, X. Sun, Z. Bie, and G. Strbac. (2023c). "Coordinating multi-energy microgrids for integrated energy system resilience: A multi-task learning approach". *IEEE Transactions on Sustainable Energy*.

Wang, Y., Z. Yan, L. Sang, L. Hong, Q. Hu, M. Shahidehpour, and Q. Xu. (2023d). "Acceleration Framework and Solution Algorithm for Distribution System Restoration Based on End-to-End Optimization Strategy". *IEEE Transactions on Power Systems*. 39(1): 429–441.

Willems, J. C. (1972). "Dissipative dynamical systems part I: General theory". *Archive for rational mechanics and analysis*. 45(5): 321–351.

Williams, R. J. (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning". *Machine learning*. 8: 229–256.

Wu, X., S. Magnússon, and M. Johansson. (2023). "Distributed safe resource allocation using barrier functions". *Automatica*. 153: 111051.

Wu, Y., Y. Ye, J. Hu, P. Zhao, L. Liu, G. Strbac, and C. Kang. (2024). "Chance Constrained MDP Formulation and Bayesian Advantage Policy Optimization for Stochastic Dynamic Optimal Power Flow". *IEEE Transactions on Power Systems.*

Xia, Y., Y. Xu, Y. Wang, S. Mondal, S. Dasgupta, A. K. Gupta, and G. M. Gupta. (2022). "A safe policy learning-based method for decentralized and economic frequency control in isolated networked-microgrid systems". *IEEE Transactions on Sustainable Energy.* 13(4): 1982–1993.

Xu, T., Y. Liang, and G. Lan. (2021). "Crpo: A new approach for safe reinforcement learning with convergence guarantee". In: *International Conference on Machine Learning.* PMLR. 11480–11491.

Xu, X., H. Hu, Y. Liu, J. Tan, H. Zhang, and H. Song. (2022). "Moving target defense of routing randomization with deep reinforcement learning against eavesdropping attack". *Digital Communications and Networks.* 8(3): 373–387.

Yan, R., Q. Xing, and Y. Xu. (2023). "Multi-Agent Safe Graph Reinforcement Learning for PV Inverters-Based Real-Time Decentralized Volt/Var Control in Zoned Distribution Networks". *IEEE Transactions on Smart Grid.* 15(1): 299–311.

Yan, Z. and Y. Xu. (2022). "A hybrid data-driven method for fast solution of security-constrained optimal power flow". *IEEE Transactions on Power Systems.* 37(6): 4365–4374.

Yang, L., G. Chen, and X. Cao. (2025). "A deep reinforcement learning-based charging scheduling approach with augmented Lagrangian for electric vehicles". *Applied Energy.* 378: 124706.

Yang, T. Y., J. Rosca, K. Narasimhan, and P. J. Ramadge. (2020). "Projection-Based Constrained Policy Optimization". In: *8th International Conference on Learning Representations (ICLR).*

Yeh, C., V. Li, R. Datta, J. Arroyo, N. Christianson, C. Zhang, Y. Chen, M. M. Hosseini, A. Golmohammadi, Y. Shi, *et al.* (2024). "SustainGym: Reinforcement learning environments for sustainable energy systems". *Advances in Neural Information Processing Systems.* 36.

Ying, D., Y. Zhang, Y. Ding, A. Koppel, and J. Lavaei. (2024). "Scalable primal-dual actor-critic method for safe multi-agent rl with general utilities". *Advances in Neural Information Processing Systems*. 36.

Yu, P., H. Zhang, and Y. Song. (2024). "Adaptive Tie-Line Power Smoothing With Renewable Generation Based on Risk-Aware Reinforcement Learning". *IEEE Transactions on Power Systems*.

Yu, P., H. Zhang, Y. Song, H. Hui, and G. Chen. (2023). "District cooling system control for providing operating reserve based on safe deep reinforcement learning". *IEEE Transactions on Power Systems*. 39(1): 40–52.

Yuan, Y., K. Dehghanpour, Z. Wang, and F. Bu. (2022). "A joint distribution system state estimation framework via deep actor-critic learning method". *IEEE Transactions on Power Systems*. 38(1): 796–806.

Yuan, Z., C. Zhao, and J. Cortés. (2024). "Reinforcement learning for distributed transient frequency control with stability and safety guarantees". *Systems & Control Letters*. 185: 105753.

Zhang, H., J. Peng, H. Tan, H. Dong, and F. Ding. (2020a). "A deep reinforcement learning-based energy management framework with Lagrangian relaxation for plug-in hybrid electric vehicle". *IEEE Transactions on Transportation Electrification*. 7(3): 1146–1160.

Zhang, J., L. Sang, Y. Xu, and H. Sun. (2024a). "Networked Multiagent-Based Safe Reinforcement Learning for Low-Carbon Demand Management in Distribution Networks". *IEEE Transactions on Sustainable Energy*.

Zhang, J., Y. Guan, L. Che, and M. Shahidehpour. (2023a). "EV charging command fast allocation approach based on deep reinforcement learning with safety modules". *IEEE Transactions on Smart Grid*. 15(1): 757–769.

Zhang, M., G. Guo, S. Magnússon, R. C. Pilawa-Podgurski, and Q. Xu. (2023b). "Data driven decentralized control of inverter based renewable energy sources using safe guaranteed multi-agent deep reinforcement learning". *IEEE Transactions on Sustainable Energy*.

Zhang, M., G. Guo, T. Zhao, and Q. Xu. (2023c). "Dnn assisted projection based deep reinforcement learning for safe control of distribution grids". *IEEE Transactions on Power Systems*.

Zhang, Q., K. Dehghanpour, Z. Wang, F. Qiu, and D. Zhao. (2020b). "Multi-agent safe policy learning for power management of networked microgrids". *IEEE Transactions on Smart Grid.* 12(2): 1048–1062.

Zhang, S., R. Jia, H. Pan, and Y. Cao. (2023d). "A safe reinforcement learning-based charging strategy for electric vehicles in residential microgrid". *Applied Energy.* 348: 121490.

Zhang, S., K. Garg, and C. Fan. (2023e). "Neural graph control barrier functions guided distributed collision-avoidance multi-agent control". In: *Conference on robot learning.* PMLR. 2373–2392.

Zhang, X., Q. Wang, X. Bi, D. Li, D. Liu, Y. Yu, and C. K. Tse. (2024b). "Mitigating cascading failure in power grids with deep reinforcement learning-based remedial actions". *Reliability Engineering & System Safety.* 250: 110242.

Zhang, Y., M. Li, Y. Chen, Y. Y. Chiang, and Y. Hua. (2023f). "A Constraint-Based Routing and Charging Methodology for Battery Electric Vehicles With Deep Reinforcement Learning". *IEEE Transactions on Smart Grid.* 14(3): 2446–2459.

Zhang, Y., J. Zhao, D. Shi, and S. Chung. (2024c). "Deep Reinforcement Learning-Enabled Adaptive Forecasting-Aided State Estimation in Distribution Systems with Multi-Source Multi-Rate Data". In: *2024 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT).* IEEE. 1–5.

Zhang, Y., J. Hu, G. Min, X. Chen, and N. Georgalas. (2024d). "Joint Charging Scheduling and Computation Offloading in EV-Assisted Edge Computing: A Safe DRL Approach". *IEEE Transactions on Mobile Computing.*

Zhao, T., J. Wang, and M. Yue. (2023). "A barrier-certificated reinforcement learning approach for enhancing power system transient stability". *IEEE Transactions on Power Systems.* 38(6): 5356–5366.

Zhou, K. and J. C. Doyle. (1998). *Essentials of robust control.* Vol. 104. Prentice hall Upper Saddle River, NJ.