Safe Reinforcement Learning for Smart Grid Control and Operations

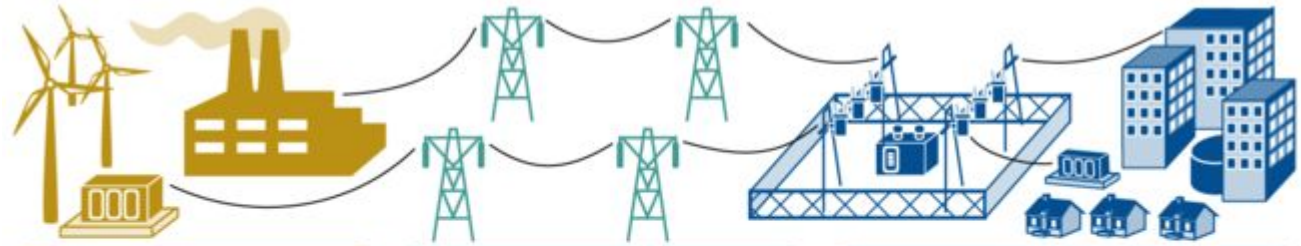
Ming Jin*
Assist. Prof. @ Virginia Tech

Vanshaj Khattar
PhD Student @ Virginia Tech

Shangding Gu
Postdoc. @ UC Berkeley

SmartGridComm Tutorial, September 17, 2024

Overview of the Modern Power Grid



Generation

Large centralized power plants)

Some utility-scale renewable plants (e.g., solar/wind farms)

Transmission

High-voltage lines

Wide-area monitoring and control systems

Distribution

Medium and low-voltage networks

Local substations and transformers

Connection to end consumers

System Operation

Balancing authorities/ISOs

Real-time grid management

Market operations

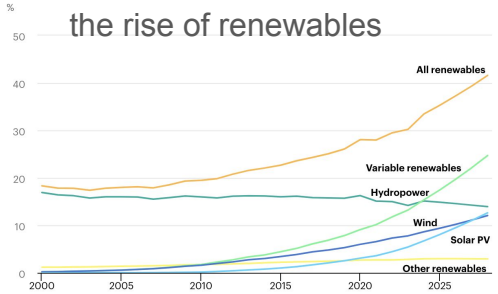
Ensure reliability and efficiency

Protective relaying	Generator Primary Control and AGC	Economic Dispatch	Unit Commitment	Maintenance scheduling	Capacity planning
Milliseconds	Seconds	Minutes	Hours	Days	Weeks	Years

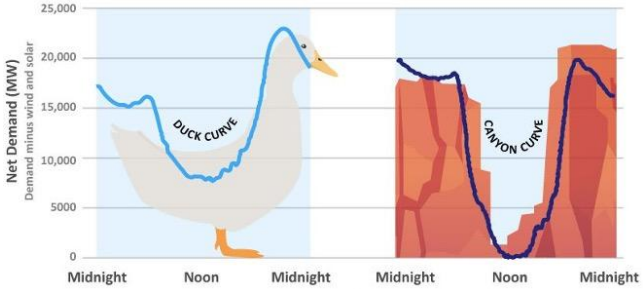
multi-scale operation

Emerging Trend (1/4): Large-scale Renewable Integration

rapid growth in **renewable energy** capacity, particularly wind and solar



the future of solar-rich grids



duck curve (2018) → canyon curve(2023)

challenges

- increased uncertainty and volatility
- reduced system inertia
- intermittency of generation

implications

- traditional dispatch methods less effective
- difficulty maintaining grid stability

opportunities

- real-time adaptive control
- data-driven forecasting
- optimal resource coordination

Emerging Trend (2/4): More Active Distribution Systems

distribution systems become more **active**, **dynamically managed** systems

500 GW by 2030:
Global Demand Response

Virtual Power Plants (VPPs)
aggregation of DERs for coordinated operation

Australia's Tesla VPP:
50,000 homes, 250 MW capacity

Key components: Distributed Energy Resources (DERs), Electric Vehicles (EVs), Demand Response (DR), Energy Storage Systems (ESS)

challenges

- ***bi***-directional power flows
- voltage regulation complexities
- coordination of DERs

implications

- need for advanced monitoring and control
- increased computation & communication requirements

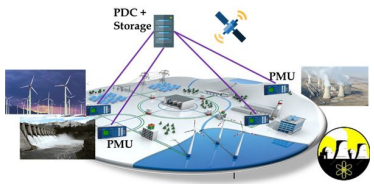
opportunities

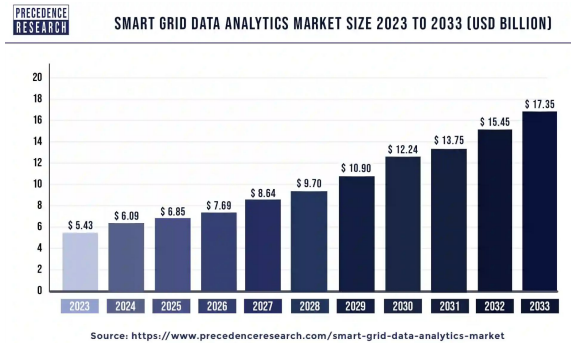
- distributed control algorithms
- real-time optimization of DERs and flexible loads

Emerging Trend (3/4): Increased Data Availability

transformation from traditional grids to **data-rich**, intelligent networks

Key components: Advanced Metering Infrastructure (AMI), Phasor Measurement Units (PMUs), Internet of Things (IoT) sensors, SCADA systems





booming *smart grid data analytics market fueled by* ***data explosion***

challenges

- *data management and processing*
- *cybersecurity and privacy concerns*

implications

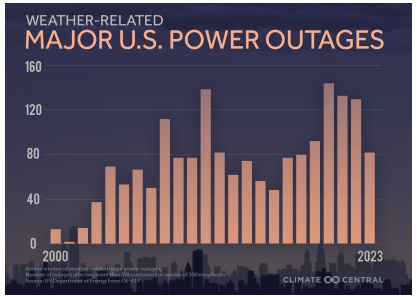
- shift from reactive to predictive grid management
- need for advanced data analytics capabilities

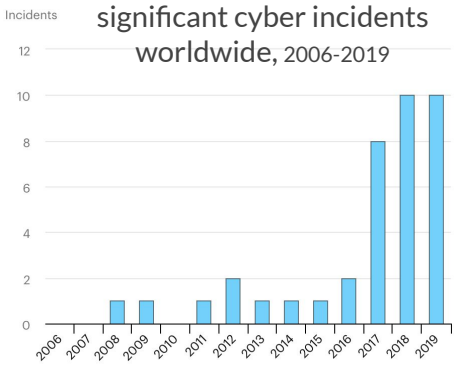
opportunities

- enhanced grid visibility and control
- predictive decisions and improved operations

Emerging Trend (4/4): Growing Focus on Grid Resilience

increasing emphasis on making the grid more **resilient** to physical and cyber disturbances





challenges

- *dealing with unpredictable disruptions*
- *protecting against evolving cyber/physical threats*

implications

- need for dynamic protection schemes
- requires rapid recovery, restoration, and self-healing

opportunities

- adaptive strategies for ***unusual*** robustness
- real-time decisions and optimal planning

Reinforcement Learning: A Powerful Tool for Modern Grids

Reinforcement Learning (RL)
an agent learns to make optimal decisions by interacting with its environment and receiving feedback

adaptive learning in dynamic environments

e.g., real-time adaptation to fluctuating renewables and changing topologies

complex multi-step planning

e.g., balancing immediate needs with multi-step constraints & performance

handling **high-dimensional** state and action spaces

e.g., coordinating multiple DERs and grid assets simultaneously

The RL Mountain Climb: Brief History

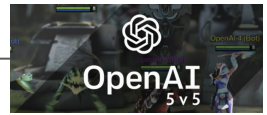
RL has evolved from simple board games to complex, real-time strategy games

…but is scaling complexity alone is NOT enough for real-world deployment

complexity

2019

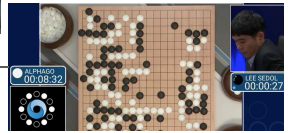
AlphaStar, achieving **grandmaster level** in **StarCraft II**



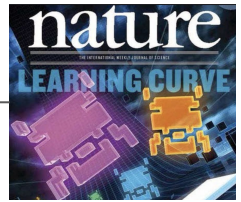
OpenAI Five in **DOTA 2**, a multiplayer strategy game

2017

2016



AlphaGo by DeepMind, defeating the world champion in Go



2013

Deep Q-Networks (DQN) by DeepMind, applying deep learning to RL and outperforming humans on several **Atari games**

1997

TD-Gammon, using temporal difference learning, achieving expert-level play

1950s

Samuel's checkers- playing program (1959)

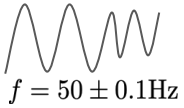
Power Systems: Safety is Paramount

simulation success

power system application

keeping frequency within ±0.1 Hz

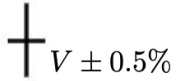
frequency regulation



$f = 50 \pm 0.1\text{Hz}$

maintaining bus voltages within ±5% of nominal

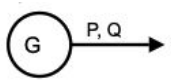
voltage control



$V \pm 0.5\%$

respecting generator, line, and voltage limits

optimal power flow



Safety Definition

maintaining system operation within technical limits, ensuring stability and reliability, while protecting equipment and human life under all conditions.

control w/o violating user comfort

demand response

ensuring stability post-disturbance

transient stability control

Overview of RL Applications in Power Systems

Optimal Power Flow (OPF)

Key Problem: Determine optimal generator outputs and voltage levels to minimize costs, while ensuring efficient operation of the grid.

Challenges: RES uncertainty, fast computation

Why RL?: Fast decisions and adaptability

Safety: Line flow limits, voltage limits, device operational limits, economic objectives

Challenges in applying safe-RL: High dimensional state-action space, balancing economic efficiency with safety constraints

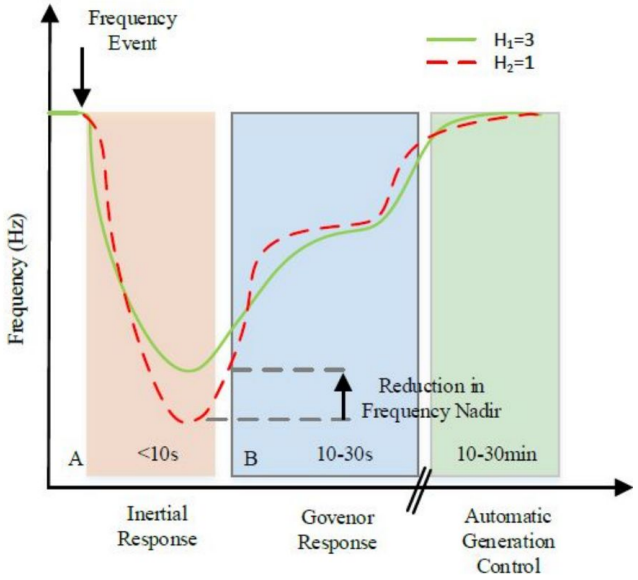
Overview of RL Applications in Power Systems

Frequency Control and Stability

Key Problem: Maintain system frequency within a narrow range around the nominal value to ensure stable operation of the grid.

Challenges: RES dynamics and uncertainty

Why RL?: Adaptability to rapid change in dynamics



(Khatibi et.al, 2019)

Safety: Frequency deviation limits, system stability, large frequency excursion risk

Challenges in applying safe-RL: safety-critical, need for robustness, balancing exploration and safety

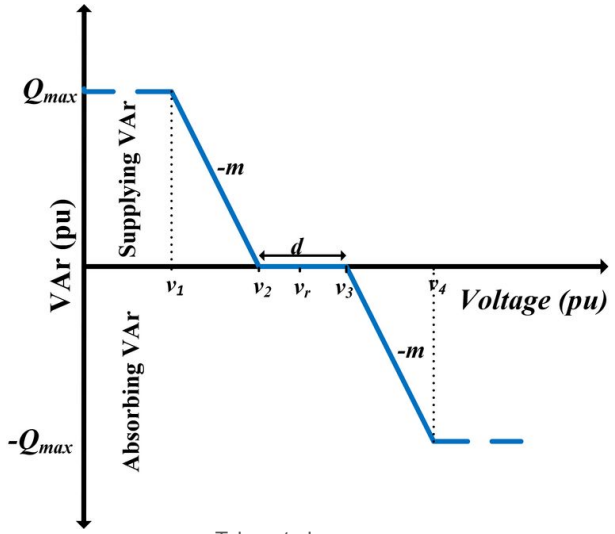
Overview of RL Applications in Power Systems

Volt-Var control (VVC)

Key Problem: Maintain acceptable voltage at all points along the distribution feeder under all loading conditions

Challenges: Rapid voltage fluctuations, reversible power flow

Why RL?: Coordinated control without full information



Teke, et al.

Safety: Voltage range, device limits, voltage profile optimization

Challenges in applying safe-RL: Coordinating multiple devices across network, ensuring safe exploration within voltage limits

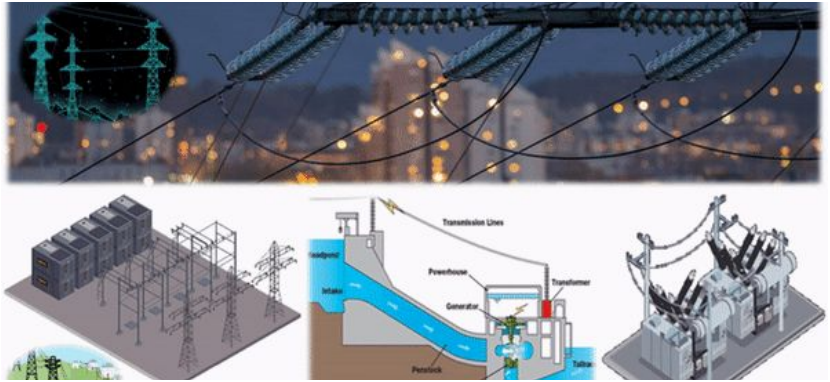
Overview of RL Applications in Power Systems

Critical load restoration (CLR)

Key Problem: Restore as many critical loads as possible using local DERs under extreme events

Challenges: DER uncertainty, topology changes, multi-step decisions

Why RL?: Fast real-time restoration, handles uncertainty



Safety: Power flow constraints, system stability, equipment protection

Challenges in applying safe-RL: dealing with changing topologies, balancing restoration speed with system stability, handling discrete restoration actions

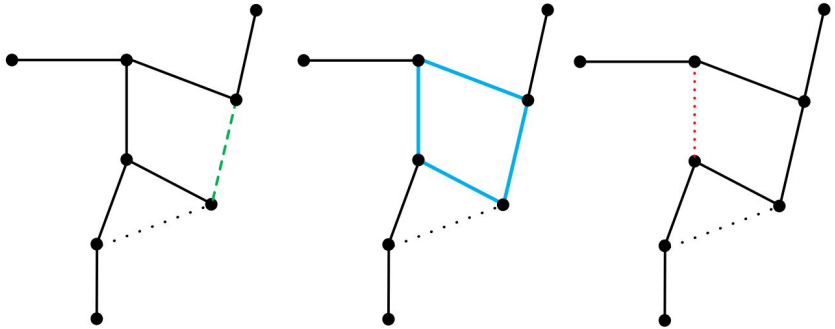
Overview of RL Applications in Power Systems

Distribution Network Reconfiguration (DNR)

Key Problem: Optimize feeder topology to satisfy certain operator's objectives

Challenges: incomplete information, computation time for big networks

Why RL?: Real-time reconfiguration, handle uncertainty



Alberto et.al

Safety: Voltage and frequency stability, radial topology constraints, total losses over time

Challenges in applying safe-RL: large discrete action space (switch operations), ensuring feasibility of actions, balancing exploration with safe operation, computational complexity of large networks

Overview of RL Applications in Power Systems

EV Charging in Microgrids

Key Problem: Optimize the charging schedules of EVs in microgrids

Challenges: variability of EV charging demands, integration of RES, grid stability

Why RL?: RL can adapt to changes in charging demands and supply

Safety: grid voltage and power limits, grid stability, user satisfaction

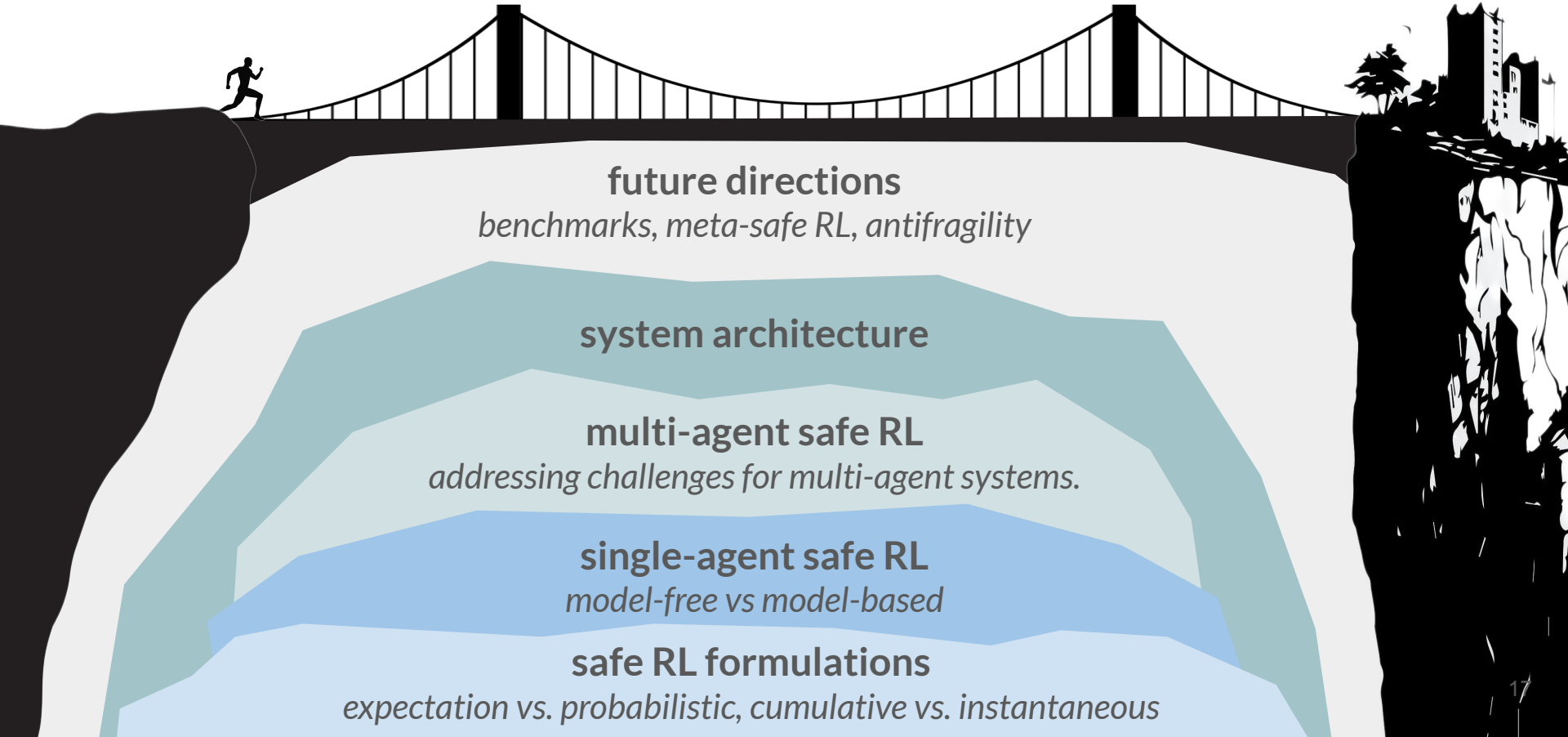
Challenges in applying safe RL: Handling high variability in EV arrival/departure time, balancing grid stability with user satisfaction, charging coordination with RES uncertainty

Key challenges of applying safe RL to power systems

Challenges

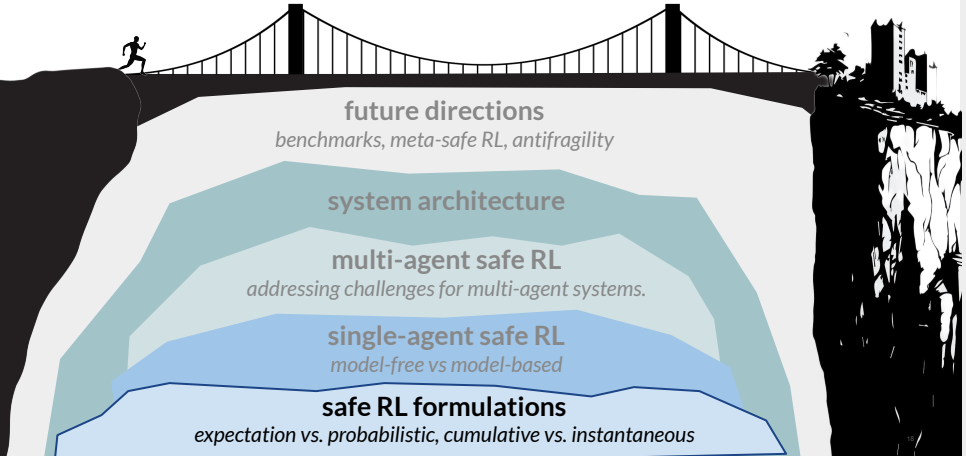
- Zero constraint violation
- Multiple constraints with conflicting objectives
- Robust performance across various operating conditions and changing constraints
- Balancing local and global optimization objectives
- Dealing with sim-to-real transfer while ensuring safety
- Computational complexity of large-scale systems

Safe RL as Bridge: Overview



safe RL formulations

presented by Ming Jin

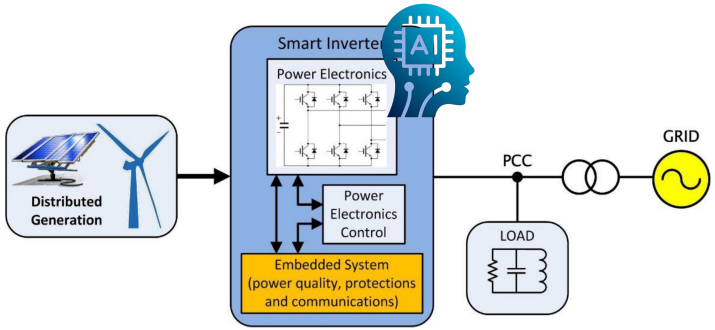


Key Objectives

- understand Constrained MDPs (CMDPs) and various safe RL formulations
- compare and contrast different safety constraint types (expectation-based vs. probability-based, instantaneous vs. cumulative)
- explore real-world applications and considerations of safe RL in energy systems

Presenter: **Ming**

Safe RL Formulations (1/6)



RL Goal

find the optimal policy that maximizes the **expected cumulative reward over time**:

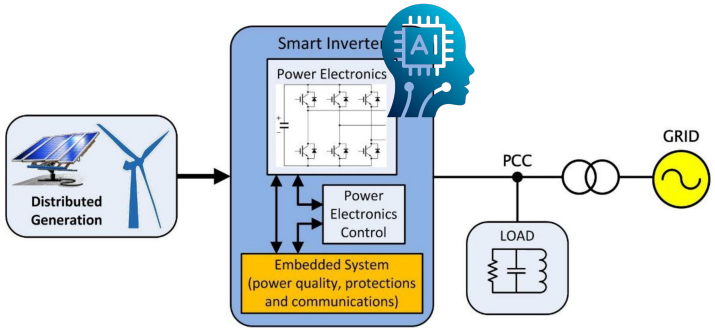
$$\max_{\pi} V_r^{\pi}(\rho) := \mathbb{E}_{\pi}\left[\sum_{t=0}^{H} \gamma^t r(s_t, a_t)\right]$$

MDPs (Markov Decision Processes):
formalizing single-agent (e.g., solar panel) decisions

$$(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, \rho)$$

- $\mathcal{S}$ **state:** the current snapshot of the system (e.g., voltage at bus)
- $\mathcal{A}$ **action:** the choices the agent can make (e.g., adjusting active/reactive power injections)
- r **reward:** the feedback that guides learning (e.g., positive for regulating the voltage)
- $\mathcal{P}$ **transition:** how actions change the state (e.g., AC power flows)
- γ **discount factor** (between 0 to 1)
- ρ **initial state distribution**

Safe RL Formulations (2/6)



Safe RL general formulation

find the optimal policy that maximizes the expected cumulative reward over time while **ensuring safety**:

$$\max_{\pi} V_r^{\pi}(\rho) := \mathbb{E}_{\pi}\left[\sum_{t=0}^{H} \gamma^t r(s_t, a_t)\right]$$

subject to **safety constraint**

$$\pi \in \Pi_{\text{safe}}$$

CMDPs (Constrained MDPs):

standard framework for Safe RL

$$(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, \rho) \cup (c, \xi)$$

c **safety cost function** (e.g., voltage limits)

ξ **safety threshold**

Expected Cumulative Cost Constraint:

Π_{safe} : all policies π such that

$$V_c^{\pi}(\rho) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{H} \gamma^t c(s_t, a_t)\right] \leq \xi$$

👍mirrors reward structure, most common
❎delayed feedback, no guarantee for trajectories

Expectation-based Constraints (3/6)

Safe RL general formulation

$$\max_{\pi} V_r^{\pi}(\rho) := \mathbb{E}_{\pi}\left[\sum_{t=0}^{H} \gamma^t r(s_t, a_t)\right]$$

subject to **safety constraint**

$$\pi \in \Pi_{\text{safe}}$$

expected cumulative cost constraint

$$V_c^{\pi}(\rho) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{H} \gamma^t c(s_t, a_t)\right] \leq \xi$$

special case:

$$c(s,a) = \mathbb{I}(s \in S_{\text{unsafe}})$$

state constraint

$$\mathbb{E}_{\pi}\left[\sum_{t=0}^{H} \gamma^t \mathbb{I}(s_t \in S_{\text{unsafe}})\right] \leq \xi$$

👍focuses specifically on avoiding unsafe states
❎still cumulative, so delayed feedback

E.g., CISR [Turchetta et al., 2020]

address the **delayed feedback** issue

expected **instantaneous** safety

$$\mathbb{E}_{\pi}\left[c(s_t, a_t)\right] \leq \xi_t, \quad \forall t \in [H]$$

👍immmediate feedback on violations
❎may be overly conservative

E.g., a grid with fluctuating renewable energy input, OptLayer [Pham et al., 2018]

Probability-based Constraints (4/6)

Markov's inequality $\mathbb{P}(X \geq a) \leq \mathbb{E}[X]/a$

*probability constraints are **stricter***

$$\mathbb{E}_\pi[c(s_t, a_t)] \leq \xi_t \Leftarrow \mathbb{P}_\pi[c(s_t, a_t) \leq \xi_t] = 1$$

*expectation as **conservative** approximation*

$$\mathbb{E}_\pi[c(s_t, a_t)] \leq \delta\xi_t \Rightarrow \mathbb{P}_\pi[c(s_t, a_t) \leq \xi_t] \geq 1 - \delta$$

expected cumulative cost constraint

$$V_c^\pi(\rho) = \mathbb{E}_\pi\left[\sum_{t=0}^{H} \gamma^t c(s_t, a_t)\right] \leq \xi$$

***Motivation:** move from average-case to worst-case safety, guarantee safety for every trajectory*

chance constraint on **cumulative** cost

$$\mathbb{P}_\pi\left[\sum_{t=0}^{H} \gamma^t c(s_t, a_t) \leq \xi\right] \geq 1 - \delta$$

👍safety for all trajectories
❎can be overly conservative, delayed feedback

E.g., Sauté RL [Sootla et al., 2022]

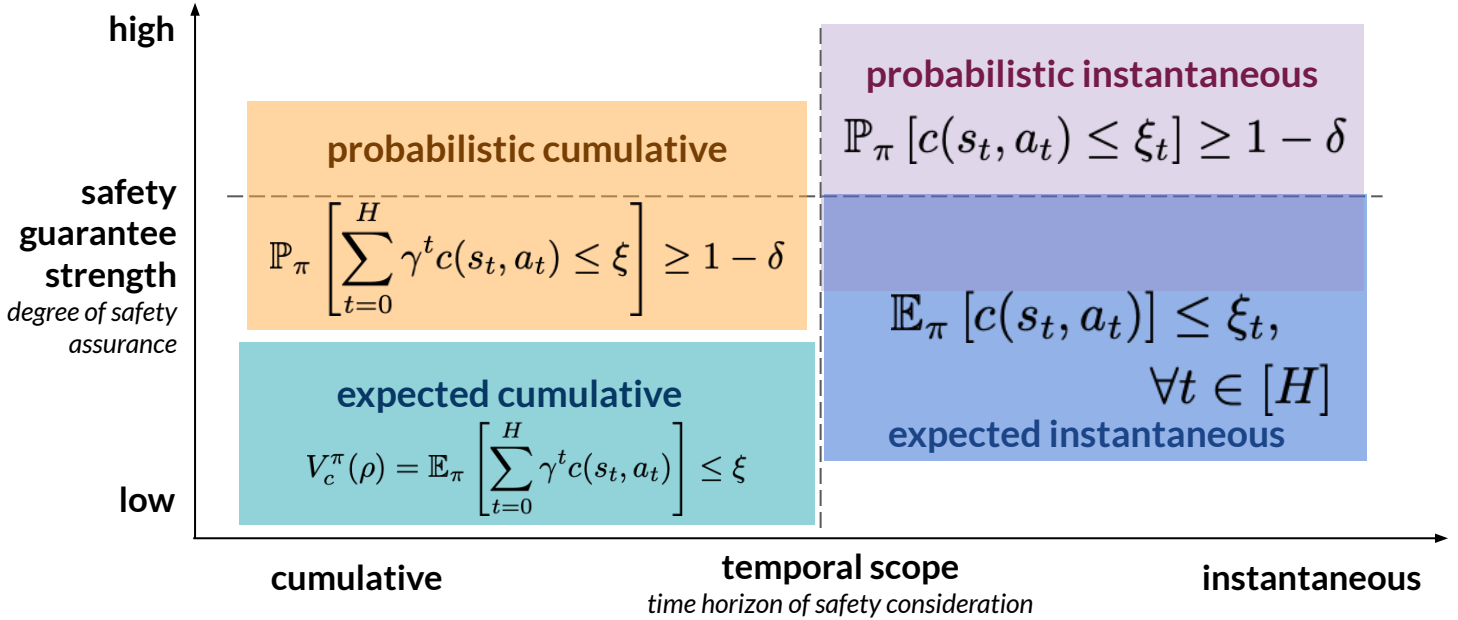
chance constraint on **immediate** cost

$$\mathbb{P}_\pi\left[c(s_t, a_t) \leq \xi_t\right] \geq 1 - \delta \quad \forall t \in [H]$$

👍immediate, strict safety guarantees
❎most conservative approach

E.g., MASE [Wachi et al., 2023]

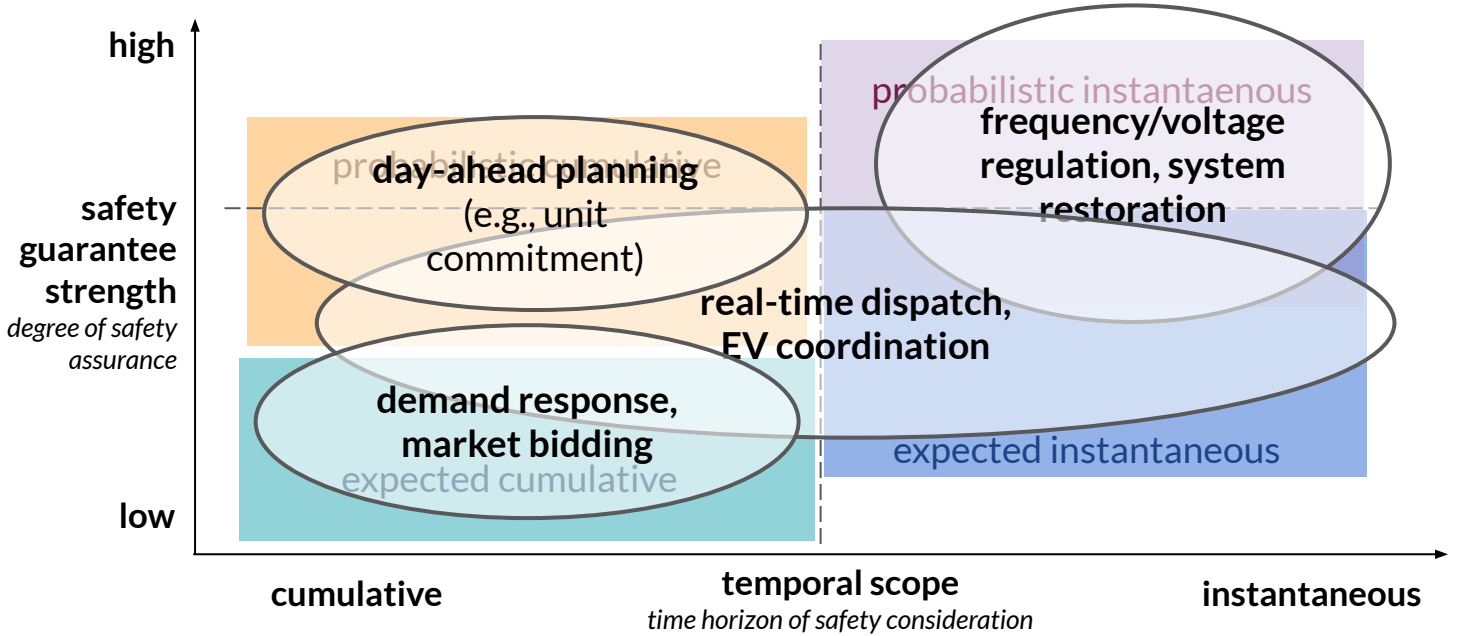
Safe RL Formulations: Comparison (5/6)



Key takeaways

- Trade-off: stronger guarantees vs. computational complexity
- instantaneous constraints: Higher safety assurance, but potentially more conservative
- Choice depends on safety criticality and resources

Real-World Applications: Energy Grid (6/6)

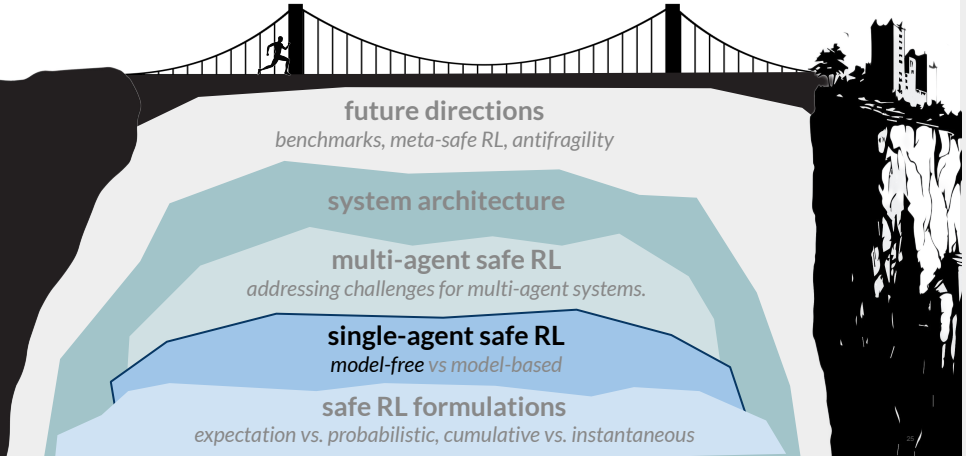


- Clear distinction between instantaneous and cumulative constraints
- Safety criticality varies by grid function

safe model-free RL

methodology

presented by Shangding Gu



Key Objectives

- understand the framework of safe model-free reinforcement learning
- know the key ideas of primal-dual and primal based safe reinforcement learning
- explore popular safe model-free reinforcement learning algorithms

Model-free methods (Shangding)

- Primal-Dual Methods [Achiam et al., 2017]
- Primal Methods [Xu et al., 2021]
- Popular model-free safe RL methods: CPO [Achiam et al., 2017], MACPO [Gu et al., 2017], CRPO [Xu et al., 2021], etc.
- For each typical method we will introduce 3 SOTA baselines

Primal-Dual methods: Ensure safety by optimizing dual parameters.

The balance between reward and cost

Primal Methods: Directly optimize reward OR safety performance.

Model-free methods

Primal-dual based Methods:

Integrate safety into reward objective via a safety weight.

$$(\pi_*, \lambda_*) = \arg\min_{\lambda} \max_{\pi} \{V_r^{\pi}(\rho) - \lambda(V_c^{\pi}(\rho) - \xi)\}$$

$$\pi_{t+1} = \arg\max_{\pi \in \Pi_0} \{V_r^{\pi} - \lambda_t(V_c^{\pi}(\pi) - \xi)\} \qquad \lambda_{t+1} = \lambda_t + \eta(\xi - V_c^{\pi}(\pi_t))_+$$

Balance reward V_r and safety V_c, searching lambda and safe policy. If safety violations happen, we adjust lambda (dual variable's weight) to consider more reward (we also optimize reward at the same time); if no safety violations, we consider more reward via adjusting lambda.

Model-free methods

Primal-dual based Methods:

$$(\pi_*, \lambda_*) = \arg\min_{\lambda} \max_{\pi} \{V_r^\pi(\rho) - \lambda(V_c^\pi(\rho) - \xi)\}$$

$$\pi_{t+1} = \arg\max_{\pi \in \Pi_0} \{V_r^\pi - \lambda_t(V_c^\pi(\pi) - \xi)\}$$

$$\lambda_{t+1} = \lambda_t + \eta(\xi - V_c^\pi(\pi_t))_+$$

Primal-Dual methods:

CPO [Achiam et al., 2017] optimize reward and cost as two surrogate models, its optimization is similar to TRPO [Schulman et al., 2015]

$$\pi_{k+1} = \arg\max_{\pi \in \Pi_\theta} J(\pi)$$
$$\text{s.t. } J_{C_i}(\pi) \leq d_i \quad i = 1, ..., m$$
$$D(\pi, \pi_k) \leq \delta.$$

surrogate models

$$\pi_{k+1} = \arg\max_{\pi \in \Pi_\theta} \underset{\substack{s \sim d^{\pi_k} \\ a \sim \pi}}{\mathrm{E}} [A^{\pi_k}(s,a)]$$
$$\text{s.t. } J_{C_i}(\pi_k) + \frac{1}{1-\gamma} \underset{\substack{s \sim d^{\pi_k} \\ a \sim \pi}}{\mathrm{E}} \left[A^{\pi_k}_{C_i}(s,a)\right] \leq d_i \quad \forall i$$
$$\bar{D}_{KL}(\pi || \pi_k) \leq \delta.$$

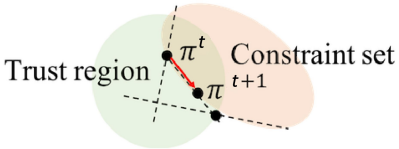
Parameterized using neural networks

$$\theta_{k+1} = \arg\max_{\theta} \; g^T(\theta - \theta_k)$$
$$\text{s.t.} \quad c_i + b_i^T(\theta - \theta_k) \leq 0 \quad i = 1, ..., m$$
$$\frac{1}{2}(\theta - \theta_k)^T H(\theta - \theta_k) \leq \delta.$$

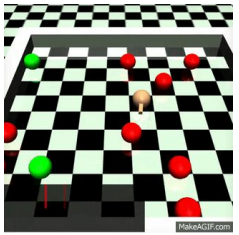
Convert it into a dual problem and search the dual parameters to optimize safety.

Primal-Dual methods:

CPO [Achiam et al., 2017] optimize reward and cost as two surrogate models, its optimization is similar to TRPO [Schulman et al., 2015]

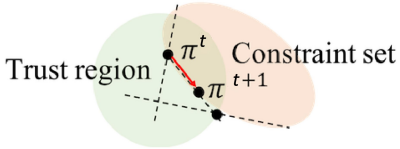


Update procedures for CPO:
CPO computes the update by simultaneously considering the trust region (light green) and the constraint set (light orange). CPO becomes infeasible when these two sets do not intersect.

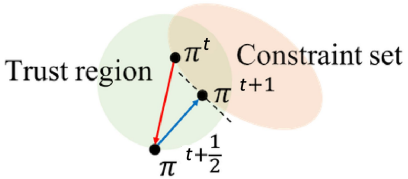


Primal-Dual methods:

PCPO [Yang et al., 2020]



Update procedures for CPO:
CPO computes the update by simultaneously considering the trust region (light green) and the constraint set (light orange). CPO becomes infeasible when these two sets do not intersect.



Update procedures for PCPO:
In step one (red arrow), PCPO follows the reward improvement direction in the trust region (light green).
In step two (blue arrow), PCPO projects the policy onto the constraint set (light orange).

Primal-Dual methods:

SAC-Lag, TRPO-Lag, and **PPO-Lag** [Ray et al., 2019] share the same key idea as follows,

```
    Returns:
        The advantage function combined with reward and cost.
    """
    penalty = self._lagrange.lagrangian_multiplier.item()
    return (adv_r - penalty * adv_c) / (1 + penalty)
```

PPO-Lagrangian method:

- easy to implement and use for safe learning;
- however, it is hard to ensure learning safety and stability.

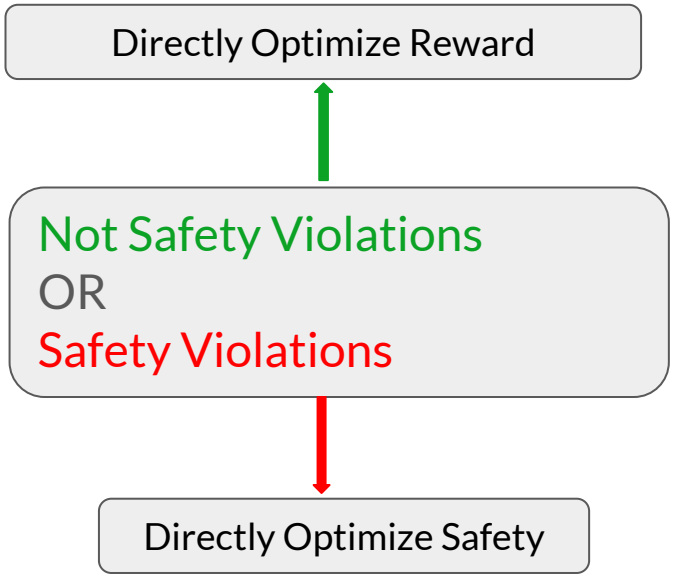
PPO-Lagrangian methods like other **primal-dual based methods**, have some issues, e.g.,

- depley safety optimization;
- sensitive to initial points;
- need to fine-tune hyper-parameters.

Primal based methods can alleviate these issues by directly optimize safety and reward performance.

Primal based Methods:

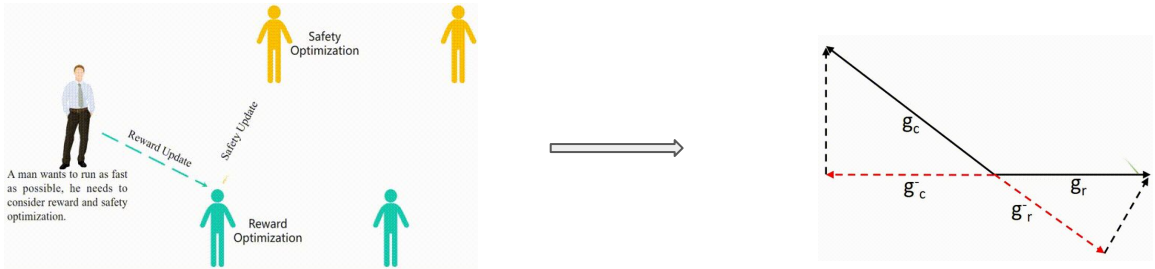
Only optimize $\max_{\pi} V_r^{\pi}$ if no safety violation happens;
Only optimize $\min_{\pi} V_c^{\pi}$ if safety violation happens.



Model-free methods

Primal methods:

CRPO: hard switching via the evaluation of safety violation [Xu et al., 2021].

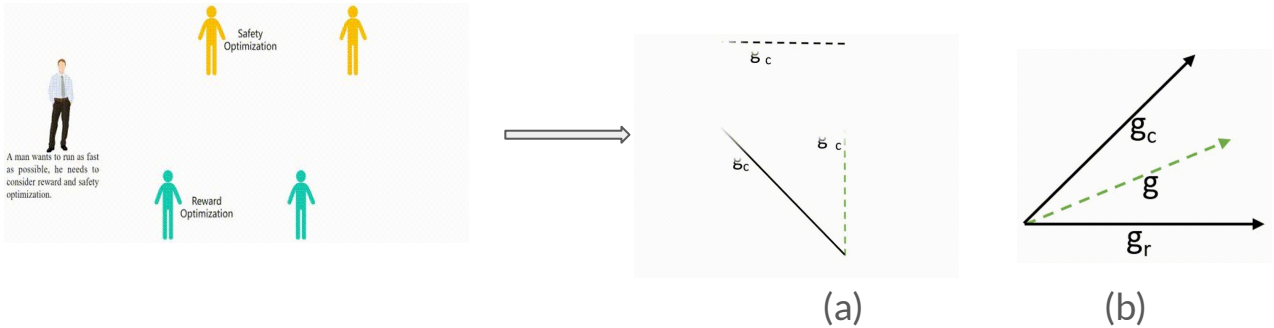


With the metric of **safety violations**, **hard switch** the gradient optimization to reward objective or safety objective.

Model-free methods

Primal methods:

PCRPO: Soft switching through gradient manipulation [Gu et al., 2024].

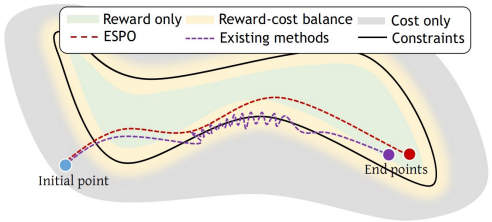


With the metric of **safety violations and gradient conflicts**, **soft switch** the gradient optimization to reward objective or safety objective.

Model-free methods

Primal methods:

ESPO: Sample manipulation with soft switching through gradient manipulation [Gu et al., 2024].



$$X_{t+1} = \begin{cases} X + X\zeta_t^+, \text{ if } \theta_{r,c} > 90°, \\ X + X\zeta_t^-, \text{ if } \theta_{r,c} \leq 90°. \end{cases}$$

With the metric of **gradient conflicts**, **dynamically adjust sample size** during three stages:

- Safety Violations
- Soft Constraint Violations
- No Violations

Summary:

Features of primal dual based methods:
1. Easy to use.
2. Can be adapted for any RL algorithms.
3. Theoretical guarantees.

Need to further investigate:
1. There is a delay to optimize safety.
2. Need to fine tune dual parameters.

Features of primal based methods:
1. It is not sensitive to the initial points.
2. Do not need to fine tune the dual parameters.
3. Theoretical guarantees.

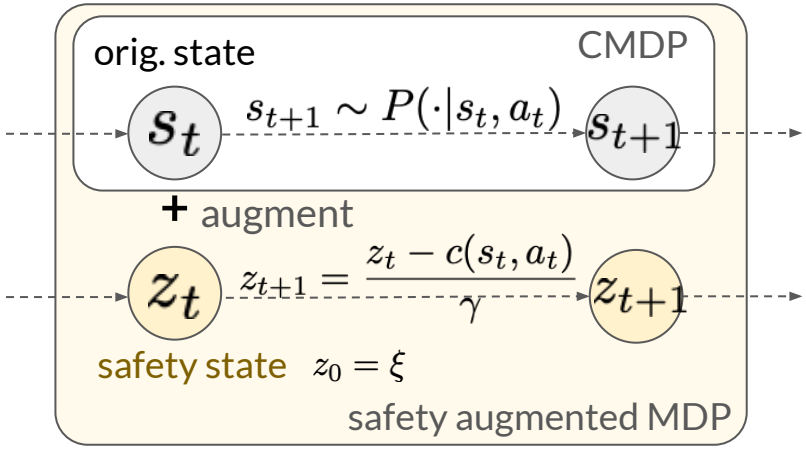
Need to further investigate:
1. For multiple constraints
2. Sample efficiency

Features of our proposed methods (PCRPO, ESPO):
1. Easy to use.
2. Can be adapted for any RL algorithms.
3. Not oscillation issues.
4. Do not need to fine-tune the dual parameters.

Need to deploy it in real-world applications.

Safety State Augmentation Technique

Augment MDP with a **safety state** so the policy directly reason about constraint satisfaction

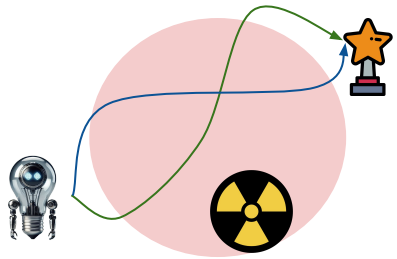


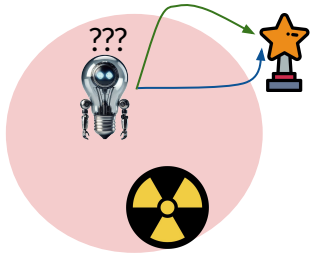
$$z_t = \gamma^{-t}\left(\xi - \sum_{t'=0}^{t} \gamma^{t'} c(s_{t'}, a_{t'})\right)$$ [Sootla et al., 2022a]

tracks the remaining safety budget

orig. CMDP $\Leftrightarrow$ safety augmented MDP

almost-sure safety [Sootla et al., 2022a]
expecrted cumulative cost safety [Sootla et al., 2022b]





[Sootla et al., 2022a]

Plug-and-Play and **Generalization**:

- compatible w/ any RL algorithm
- policies trained for one budget d generalize to others by adjusting

Other augmentations: CVaR constraints [Chow et al., 2017], Lagrangian multiplier [Calvo-Fullana et al., 2021]

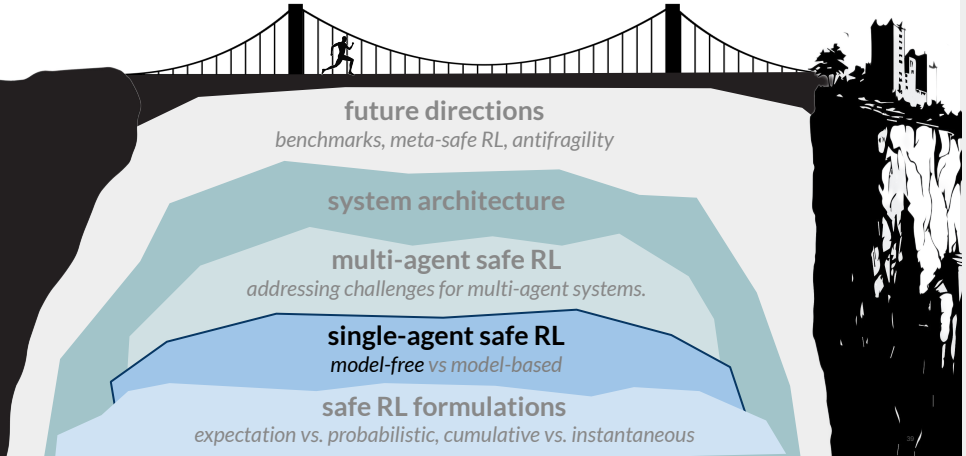
"Simmering" safety budget [Sootla et al., 2022b]: lower budgets early on can reduce safety cost spikes; gradually increase budget to learn safe behavior first

also applied to safe multi-agent RL [Wang et al. 2023]

safe model-free RL

case studies

presented by Vanshaj Khattar



Key Objectives

- understand the framework of safe model-free/model-based reinforcement learning
- know the key ideas of primal-dual and primal based safe reinforcement learning
- explore popular safe model-free reinforcement learning algorithms

Safe RL applications in EV control

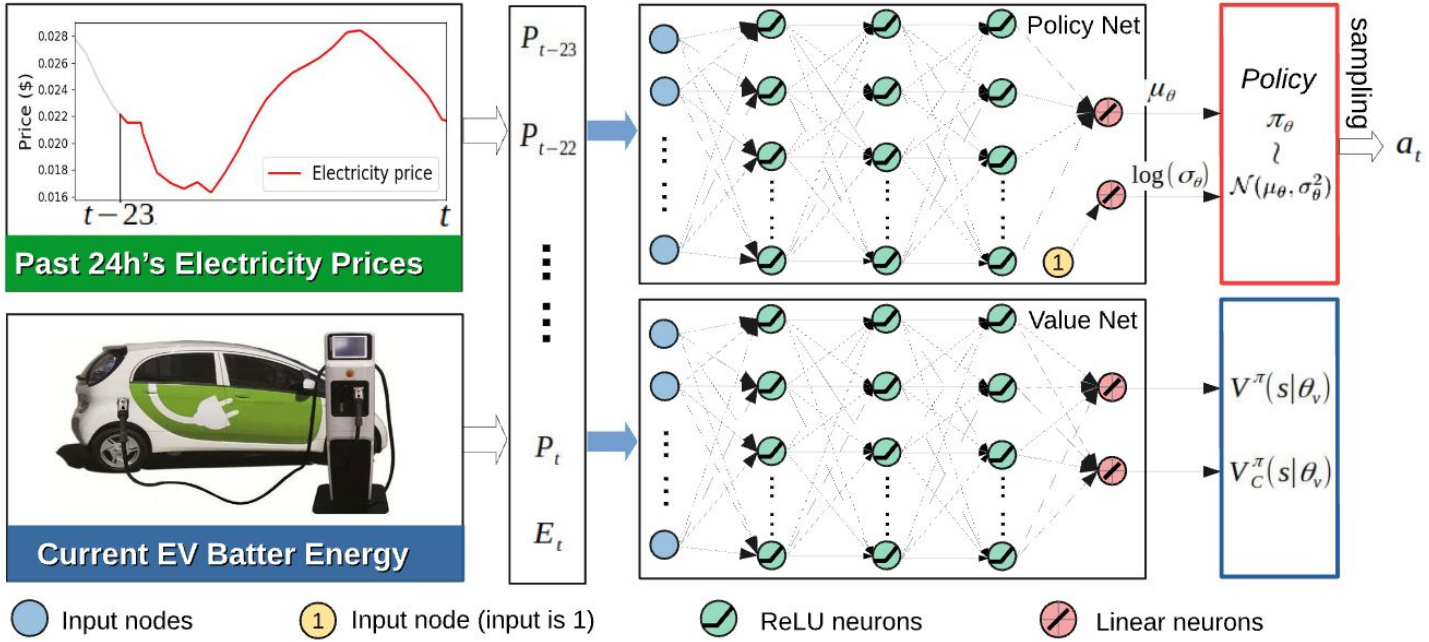
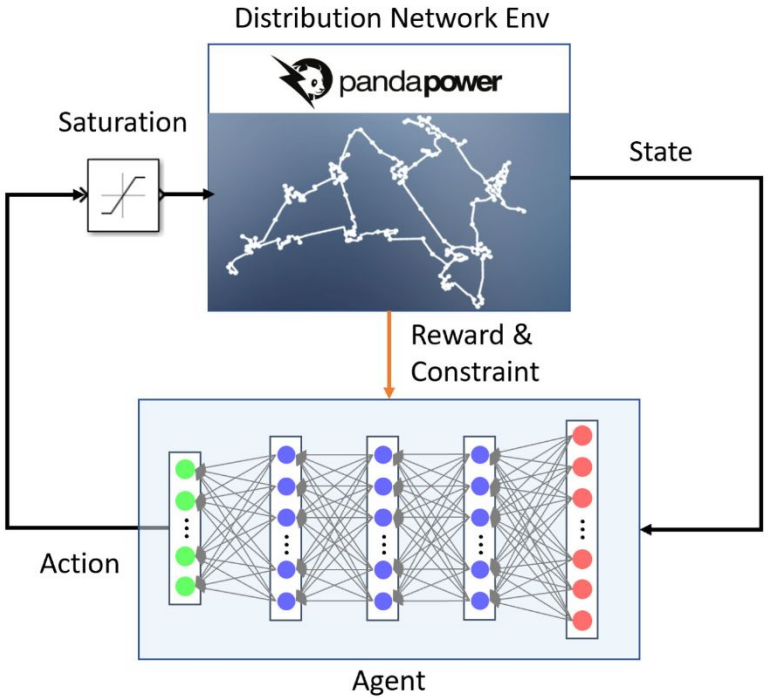


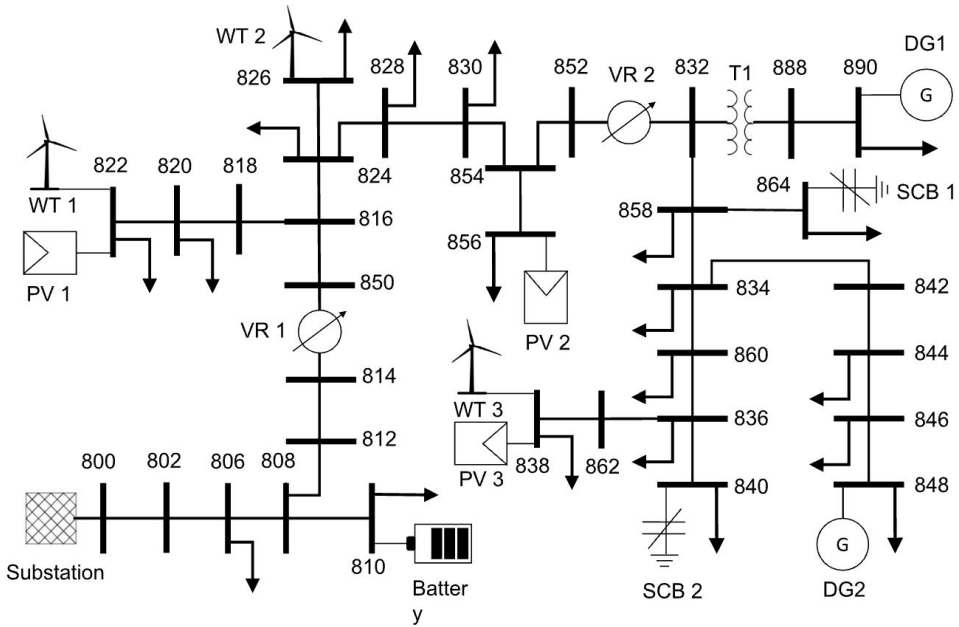
Fig. 1. Overall architecture of the designed policy network and the value network. The inputs of the both networks are the system state, i.e., the past 24h's electricity prices $P_{t-23}, \ldots, P_t$ and the current EV battery energy E_t. The policy network extracts features from the state information and outputs the mean values μ_θ and logarithmic standard deviations $\log(\sigma_\theta)$ of a normal distribution. By sampling from the normal distribution, the policy π_θ generates a charging/discharging action a_t for the EV. The value network shares the same architecture with the policy network. It extracts features from the state information and outputs the state-values $V^\pi(s|\theta_v)$ and $V_C^\pi(s|\theta_v)$.

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

Safe RL applications in EV control



Digraph of the overall training scheme. The saturation block limits out-of-range actions to their upper or lower bounds.



Modified IEEE-34 node test feeder system

Li, H., & He, H. (2022). Learning to operate distribution networks with safe deep reinforcement learning. IEEE Transactions on Smart Grid, 13(3), 1860-1872.

Application: Real-time AC-OPF

Goal: Solve real-time AC-OPF

Constraints: violation of active/reactive power constraints, voltage amplitude and power flow constraints

A case study on using **safe deep-RL** method based on **Lagrange advantage function**

- set of generators: $\mathcal{G}$
- cost coefficients: c_{1i}, c_{2i}, c_{0i}
- generator active power bus i: P_{gi}
- generator reactive power bus i: Q_{gi}
- voltage phase angle difference bus i and j : δ_{ij}
- set of buses: $\mathcal{N}$
- Non-zero net load buses: $\tilde{\mathcal{N}} \subset \mathcal{N}$

Optimization formulation

$$\min \sum_{i \in \mathcal{G}} \left(c_{2i} P_{gi}^2 + c_{1i} P_{gi} + c_{0i} \right),$$

under power flow and operational constraints

Wu, Pengfei, et al

Application: Real-time AC-OPF

CMDP formulation:

State: $\left(P_i^t, Q_i^t, \forall i \in \tilde{\mathcal{N}}, P_{gi}^{t-1}, V_{gi}^{t-1}, \forall i \in \mathcal{G}, \sum_{i \in \tilde{\mathcal{N}}} P_i^{t+1} \right)$

active and reactive load bus i time t

Action: $\left(\Delta P_{gi}^t, \Delta V_{gi}^t, \forall i \in \mathcal{G} \right)$ → Adjustments of active output and voltage set-points

Reward: $-\sum_{i \in \mathcal{G}} \left(c_{2i} P_{gi}^2 + c_{1i} P_{gi} + c_{0i} \right)$

Cost function: $\left[C_{pg},\ C_{qg},\ C_v,\ C_{flow} \right]$ → **Constraint violations** of active/reactive power, voltage amplitude and power flow

Primal Dual PPO (PD-PPO):

$$L(\pi_\theta, \boldsymbol{\lambda}) = R(\pi_\theta) - \sum_i^m \lambda_i \left(C_i(\pi_\theta) - d_i \right),$$

$$\left(\pi^*, \boldsymbol{\lambda}^* \right) = \arg \min_{\lambda_i \geq 0} \max_\theta L(\pi_\theta, \boldsymbol{\lambda}),$$

PD-PPO

1. **Generalized advantage estimation**
2. **Importance sampling**
3. **Gradient clipping**

Application: Real-time AC-OPF

Generalized advantage estimation (GAE):

$$\hat{A}_{R,t}^{(k)} = \sum_{l=0}^{k-1} \gamma^l \delta_{t+l}^R,$$

$$\delta_t^R = r_t + \gamma V_\pi^R(s_{t+1}) - V_\pi^R(s_t),$$

$$A_{R,t}^{GAE} = (1-\lambda_{GAE})\left(\hat{A}_{R,t}^{(1)} + \lambda_{GAE}\hat{A}_{R,t}^{(2)} + \lambda_{GAE}^2 \hat{A}_{R,t}^{(3)} + \cdots\right)$$

GAE

Lagrange advantage function

$$A_{L,t}^{GAE} = A_{R,t}^{GAE} - \sum_i^m \lambda_i A_{C_i,t}^{GAE}$$

Knowledge-driven action masking (PDO-AMG): → identify crucial exploration directions

$$\begin{bmatrix} \Delta \boldsymbol{P} \\ \Delta \boldsymbol{Q} \end{bmatrix} = \begin{bmatrix} \boldsymbol{H} & \boldsymbol{N} \\ \boldsymbol{T} & \boldsymbol{L} \end{bmatrix} \begin{bmatrix} \Delta \boldsymbol{\theta} \\ \Delta \boldsymbol{V} \end{bmatrix}$$

$$M_i = \sum_{j \in \mathcal{C}_{vio}} |w_{j,i}| \Big/ \max_{1 \le k \le |\mathcal{A}|} \left(\sum_{j \in \mathcal{C}_{vio}} |w_{j,k}| \right)$$

$$\widehat{a}_i = (a_i - \mu_i) \cdot M_i + \mu_i,$$

$$p(\widehat{\boldsymbol{a}} \mid \boldsymbol{s}) = \prod_{i=1}^{2|\mathcal{G}|} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-M_i^2 (a_i - \mu_i)^2 / 2\sigma_i^2\right)$$

$H_{ij} = \partial P_i / \partial \theta_j \quad N_{ij} = \partial P_i / \partial V_j$

$L_{ij} = \partial Q_i / \partial V_j \quad T_{ij} = \partial Q_i / \partial \theta_j$

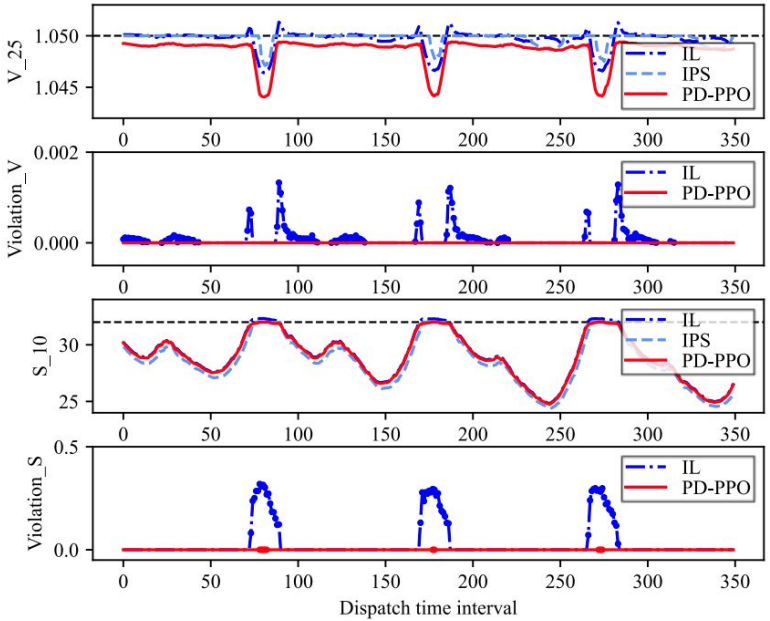
$\boldsymbol{M} = \left[M_1, \ldots, M_i, \ldots, M_{2|\mathcal{G}|} \right]$

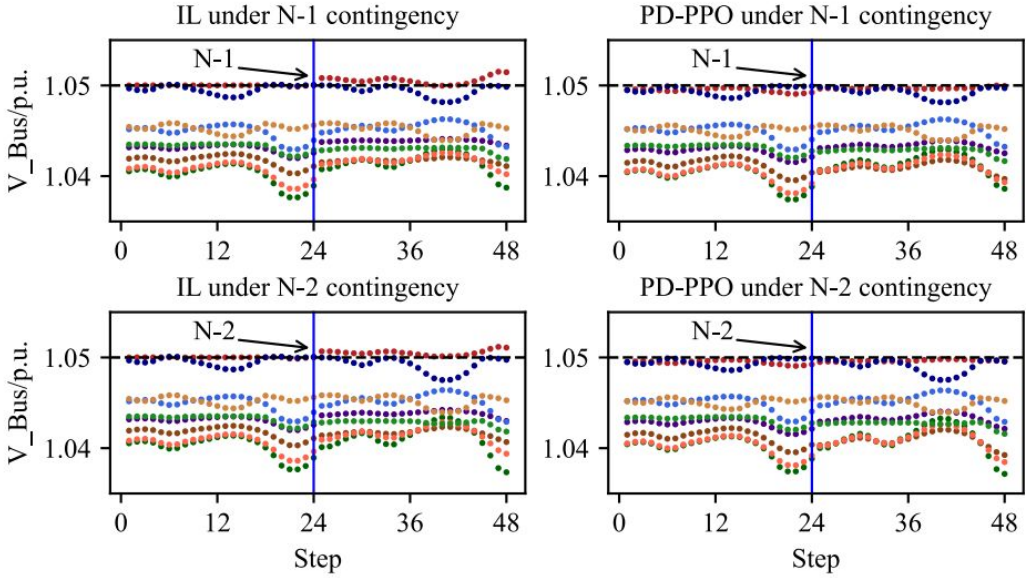
$\widehat{a}_i$ → ith modified action

Application: Real-time AC-OPF

Case study

- Case study on the IEEE 30-bus system

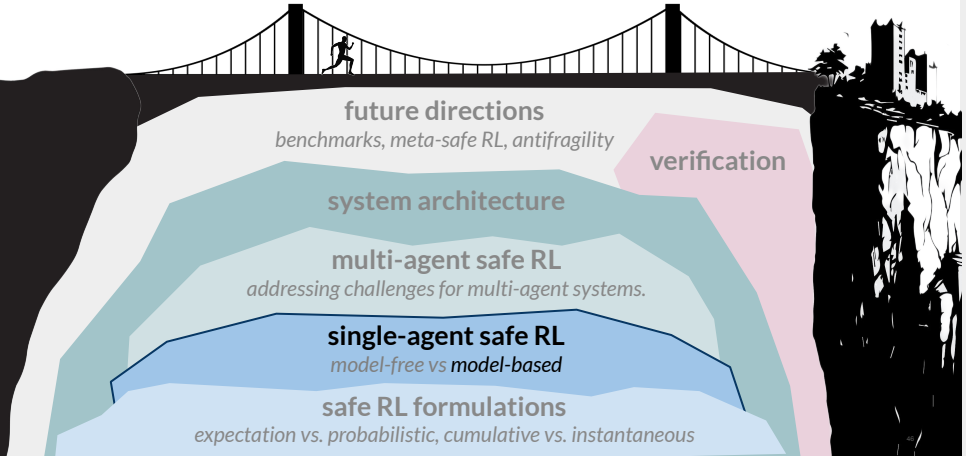




safe model-based RL

methodology

presented by Ming Jin



Key Objectives

- understand the framework of safe model-based reinforcement learning
- explore control theoretic certificates for safe MBRL, including Lyapunov functions, barrier functions, and contraction metrics

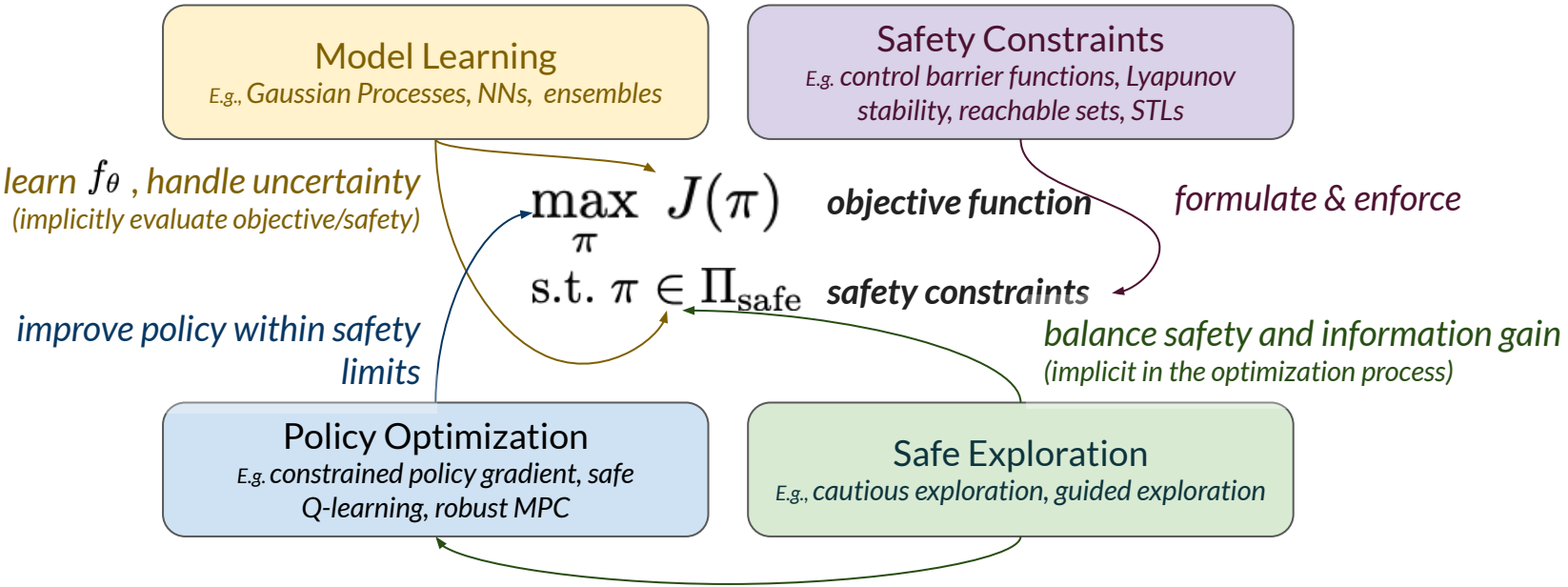
Overview of Safe Model-based RL (MBRL)

Safe MBRL: An approach combining model learning, RL, and safety constraints

model: $s_{t+1} = f_\theta(s_t, a_t, w_t)$

system dynamics

uncertainty/disturbances



Overview of Control Theoretic Notions

system dynamics (deterministic, fully observable): $s_{t+1} = f(s_t, a_t)$ control action: $a_t = \pi(s_t)$

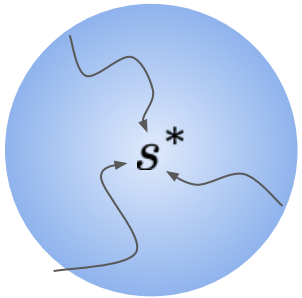
Lyapunov functions

♠ positive definite: $W(s) > 0$ for all $s \neq s^*, W(s^*) = 0$

♠ "dissipates energy": for $s \neq s^*$

$\Delta W(s) = W(f(s, \pi(s))) - W(s) < 0$

ensure system converges to a **desired equilibrium** s^*



Barrier functions

♠ safe set: $S_{\text{safe}} = \{s | h(s) \leq 0\}$

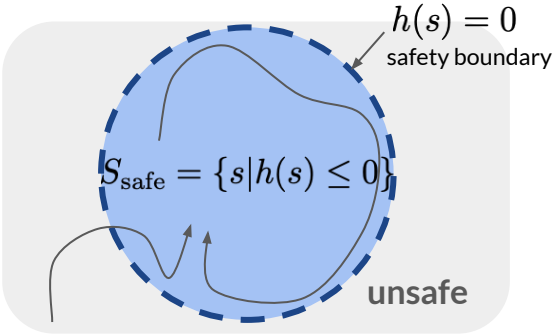
♠ "barrier" condition:

$h(f(s, \pi(s))) - h(s) \leq -\alpha(h(s))$

where $\alpha(\cdot)$ is a class K function (continuous, strictly increasing with $\alpha(0) = 0$)

(Ames, et al., 2019)

ensures system **stay within safe region** & keeps away from unsafe region at each step



Contraction metrics

♠ continuous, positive definite:

$M : \mathbb{R}^n \to \mathbb{R}^{n \times n}$

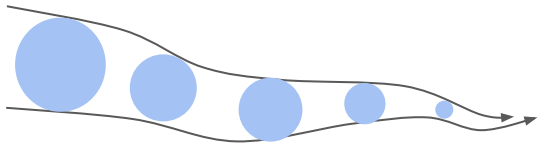
♠ "contraction" condition: for $0 < \beta < 1$,

$\delta s_{t+1}^\top M(s_{t+1}) \delta s_{t+1} \leq \beta \delta s_t^\top M(s_t) \delta s_t$

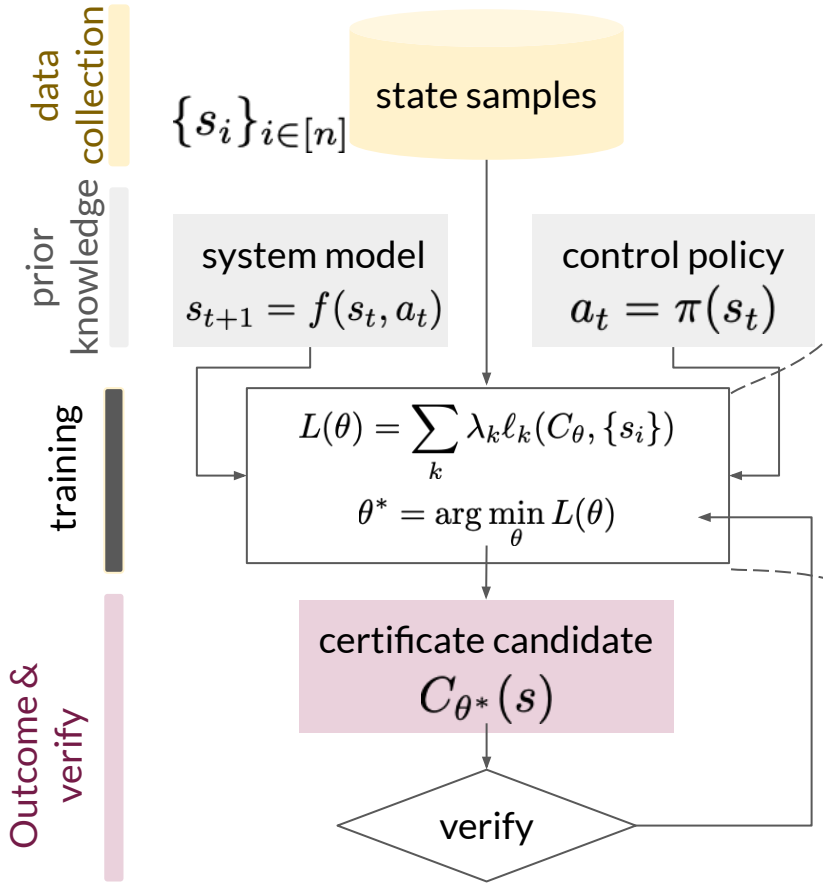
where δs is a virtual displacement between two nearby trajectories.

(Lohmiller & Slotine,, 1998)

guarantees exponential **convergence of nearby trajectories**



Neural Certificate for Control Policy



How to design loss function?

$$L(\theta) = \sum_k \lambda_k \ell_k(C_\theta, \{s_i\})$$

Key insight: convert conditions to loss

- inequality $a \le b$ → loss $[a-b]_+ := \max(0, a-b)$ or $([a-b]_+)^2$
- equality $a = b$ becomes loss term $(a-b)^2$

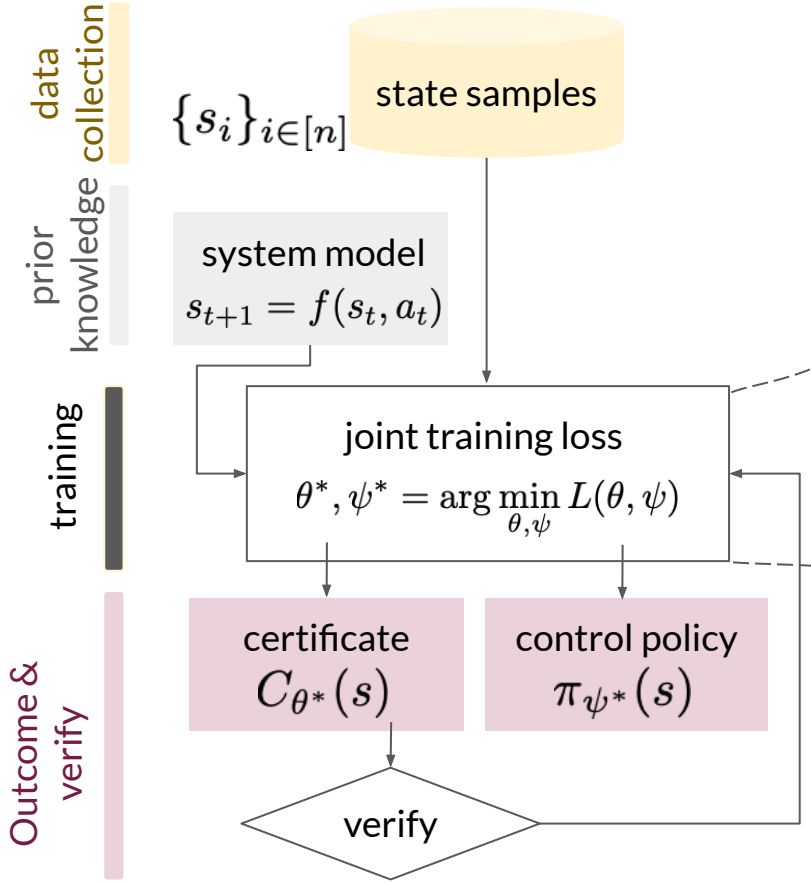
E.g., Learning-based neural Lyapunov loss

$$W(s^*) = 0 \qquad W(s) \ge 0 \qquad W(f(s,\pi(s))) - W(s) < 0$$

$$\lambda_1 W_\theta(s^*) + \lambda_2 \sum_i [-W_\theta(s_i)]_+ + \lambda_3 [W_\theta(f(s_i, a_i)) - W_\theta(s_i)]_+$$

Key Considerations:
1) sample coverage (focus on critical regions and safe set boundaries), **2)** balance loss terms (adjust weights λ_k for different conditions), **3)** add small margin for robustness, **4)** online update during execution [Qin et al, 2020]

Joint Neural Certificate and Policy Learning



How to design **joint** loss function?

$$L(\theta, \psi) = L_{\text{cer}}(\theta, \psi) + \lambda L_{\text{policy}}(\psi)$$

$L_{\text{cer}}(\theta, \psi)$: same as before, but with a_t replaced by $\pi_\psi(s_t)$

$$L_{\text{policy}}(\psi) = -V(\pi_\psi)$$

$$:= -\mathbb{E}\left[\sum_t \gamma^t r(s_t, \pi_\psi(s_t))\right]$$

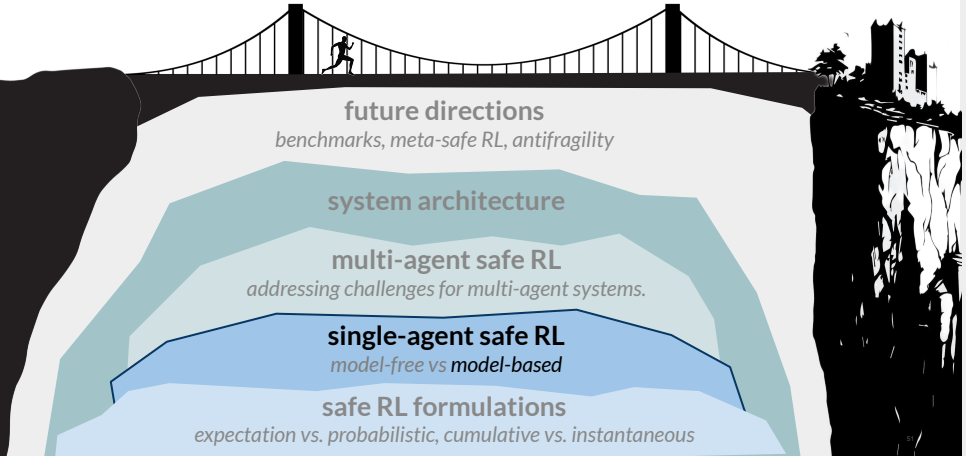
- ♠ Optimize θ and ψ simultaneously
- ♠ Verify: learner-falsifier framework (Chang et al., 2019)

Neural Lyapunov functions: (Richards et al., 2018) (Chang et al., 2019) (Choi et al., 2020) (Taylor et al., 2019); **Learning barrier functions:** (Srinivasan et al., 2020), (Qin et al., 2021); **Learning contraction metrics:** (Singh et al., 2020); **Verification:**(Abate et al., 2021) (Yang et al., 2023); **Survey:** (Dawson et al., 2022)

safe model-based RL

case studies

presented by Vanshaj Khattar



Key Objectives

- understand the framework of safe model-based reinforcement learning
- explore control theoretic certificates for safe MBRL, including Lyapunov functions, barrier functions, and contraction metrics

Application: Transient frequency control

Goal: Design controllers to maintain nodal frequencies within safety bounds during transients

Safety constraints: Deviations from the nominal frequency (e.g., 50 or 60 Hz) can lead to **equipment damage, instability, and blackouts.**

A case study using model-based safe RL with **Lyapunov stability theory** for controller design ensures **system stability** and **transient frequency safety.**

- frequency: ω_i
- active power injection: p_i
- controlled active power injection: u_i
- voltage angle difference: λ
- $\omega^\infty \triangleq \frac{\sum_{i=1}^n p_i}{\sum_{i=1}^n D_i}$

Problem formulation

$$\min_u \int_{t=0}^{T} \left\{\gamma\|\omega(t)-\omega^\infty 1_n\|^2 + \|u(t)\|^2\right\} dt$$

$$\text{s.t.} \quad \dot{\lambda} = B^T\omega$$

$$M\dot{\omega} = -D\omega - BY_b \sin\lambda + u + p,$$

$$\lim_{t\to\infty}(\lambda,\omega) = (\lambda^\infty, \omega^\infty 1_n) \quad \underline{\omega}_i \leq \omega_i \leq \bar{\omega}_i$$

inertia and damping coefficients: M_i, D_i, incidence matrix: B, time horizon: T, cost coefficient: γ nodes adjacency matrix: Y_b

Yuan, Zhenyi, et.al.

Application: Transient frequency control

Search Space of control policies

$\omega_i(t) \in [\underline{\omega}_i, \overline{\omega}_i]$ Let $\mathcal{Q}_i \triangleq \{x \mid \underline{\omega}_i \le \omega_i \le \overline{\omega}_i\}$ → How to have this set forward invariant?

Introduce a dead zone $[\underline{\omega}_i^{\text{th}}, \overline{\omega}_i^{\text{th}}]$

Sufficient condition for frequency invariance

Let $\underline{\omega}_i < \underline{\omega}_i^{\text{th}} < \overline{\omega}_i^{\text{th}} < \overline{\omega}_i$, and $\overline{\alpha}_i(\cdot)$ and $\underline{\alpha}_i(\cdot)$ be class- K functions. If for all x and p

$$\begin{cases} u_i(x,p) \le \dfrac{\overline{\alpha}_i(\overline{\omega}_i - \omega_i)}{\omega_i - \overline{\omega}_i^{\text{th}}} + q_i(x,p), & \omega_i > \overline{\omega}_i^{\text{th}}, \\ u_i(x,p) \ge \dfrac{\underline{\alpha}_i(\underline{\omega}_i - \omega_i)}{\underline{\omega}_i^{\text{th}} - \omega_i} + q_i(x,p), & \omega_i < \underline{\omega}_i^{\text{th}}, \end{cases}$$

→ $\mathcal{Q}_i$ is a forward invariant set

$$q_i(x,p) \triangleq D_i\omega_i + [BY_b]_i \sin\lambda - p_i$$

$$a(\lambda_j) \triangleq \cos\lambda_j^{\infty} - \cos\lambda_j - \lambda_j \sin\lambda_j^{\infty} + \lambda_j^{\infty} \sin\lambda_j^{\infty}$$

Constraint ensuring asymptotic stability

$$V(\lambda,\omega) \triangleq \frac{1}{2}\sum_{i=1}^{n} M_i(\omega_i - \omega^{\infty})^2 + \sum_{j=1}^{m} [Y_b]_{j,j}\, a(\lambda_j)$$

→ How to ensure $\dot{V}(\lambda,\omega) \le 0$?

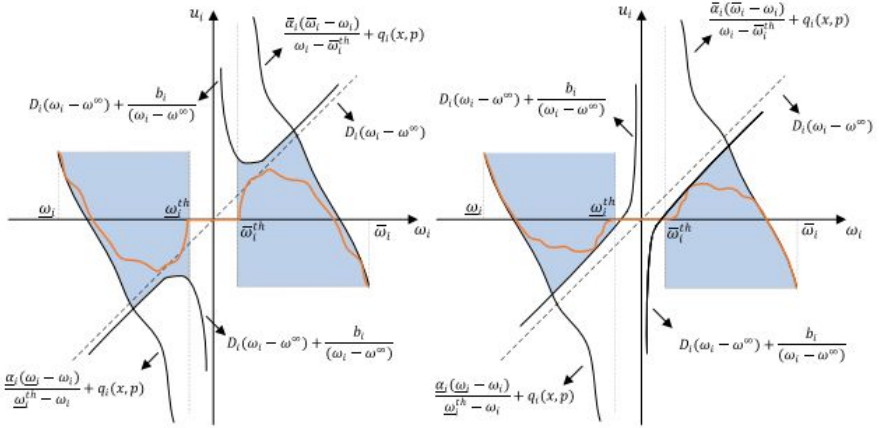
Application: Transient frequency control

Distributed, stable and safe control policies

$$\begin{cases} u_i(x,p) \le \min\left\{ \tilde{D}_i(\omega_i - \omega^\infty) + \dfrac{b_i}{(\omega_i - \omega^\infty)}, \dfrac{\overline{\alpha}_i(\overline{\omega}_i - \omega_i)}{\omega_i - \overline{\omega}_i{}^{\text{th}}} + q_i(x,p) \right\} & \omega_i > \overline{\omega}_i{}^{\text{th}}, \\ u_i(x,p) = 0 & \underline{\omega}_i{}^{\text{th}} \leqslant \omega_i \leqslant \overline{\omega}_i{}^{\text{th}}, \\ u_i(x,p) \ge \max\left\{ \tilde{D}_i(\omega_i - \omega^\infty) + \dfrac{b_i}{(\omega_i - \omega^\infty)}, \dfrac{\underline{\alpha}_i(\underline{\omega}_i - \omega_i)}{\underline{\omega}_i{}^{\text{th}} - \omega_i} + q_i(x,p) \right\} & \omega_i < \underline{\omega}_i{}^{\text{th}}. \end{cases}$$

$$0 < \tilde{D}_i < D_i$$
$$\{b_i\}_{i \in \mathcal{I}} \in \mathbb{R}^n$$

Search space of safe control policies



(a) $b_i > 0$ (b) $b_i < 0$

Application: Transient frequency control

Synthesis of distributed neural network controllers

Select class-K functions and frequency thresholds

$$u_{i,\phi}(x,p) = \frac{\sigma(\omega_i - \overline{\omega}_i^{\text{th}})}{\omega_i - \overline{\omega}_i^{\text{th}}} \left[f_i(\omega_i) - \sigma(f_i(\omega_i) - \overline{u}_{i,\phi}(x,p)) \right] + \frac{\sigma(\underline{\omega}_i^{\text{th}} - \omega_i)}{\underline{\omega}_i^{\text{th}} - \omega_i} \left[f_i(\omega_i) + \sigma(\underline{u}_{i,\phi}(x,p) - f_i(\omega_i)) \right]$$

Learning optimal control policy using RNN

First-order Euler discretization with step-size Δt

$$\min_{\phi} \frac{1}{K} \sum_{k=0}^{K-1} \gamma \|\omega(k) - \omega^{\infty} \mathbf{1}_n\|^2 + \|u_\phi(k)\|^2$$

$$\text{s.t. } \lambda(k) = \lambda(k-1) + B^{\top} \omega(k-1) \Delta t$$

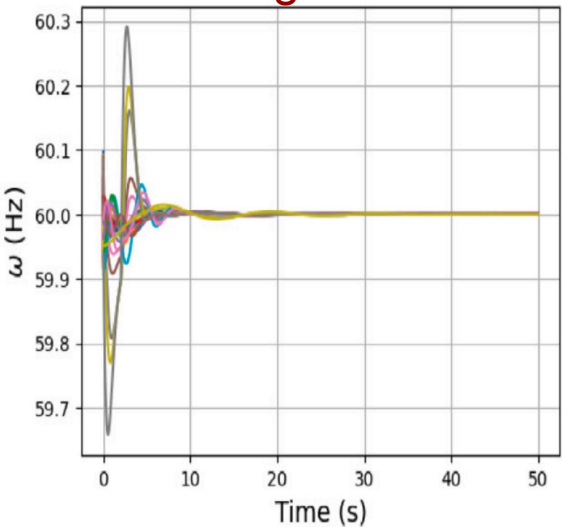
$$M(\omega(k) - \omega(k-1)) = \big[-D\omega(k-1) - BY_b \sin \lambda(k-1) + u_\phi(k-1) + p \big] \Delta t.$$

Application: Transient frequency control

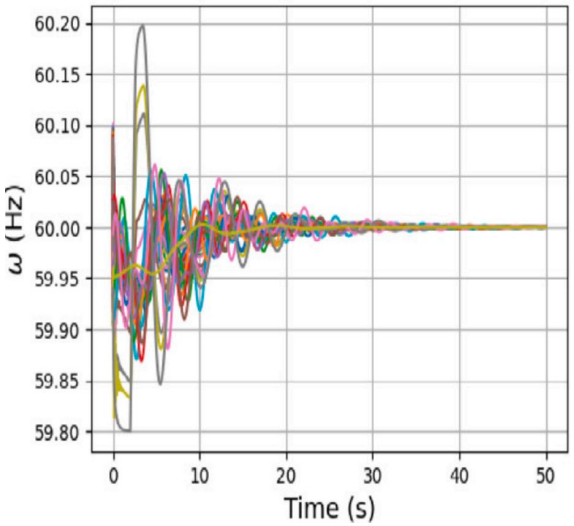
Case study on IEEE-39 bus power network

- T = 50 sec with bus 38 encountering a sudden change in power injection from (0,2] seconds.
- Nominal frequency = 60 Hz and safe region [59.8, 60.2]

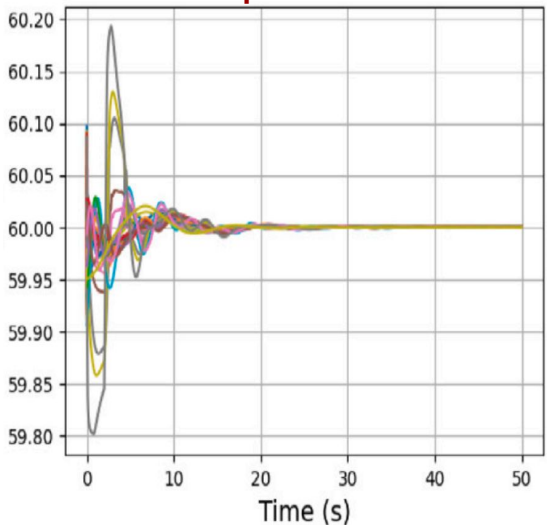
Zhang et.al.



Cui et.al.



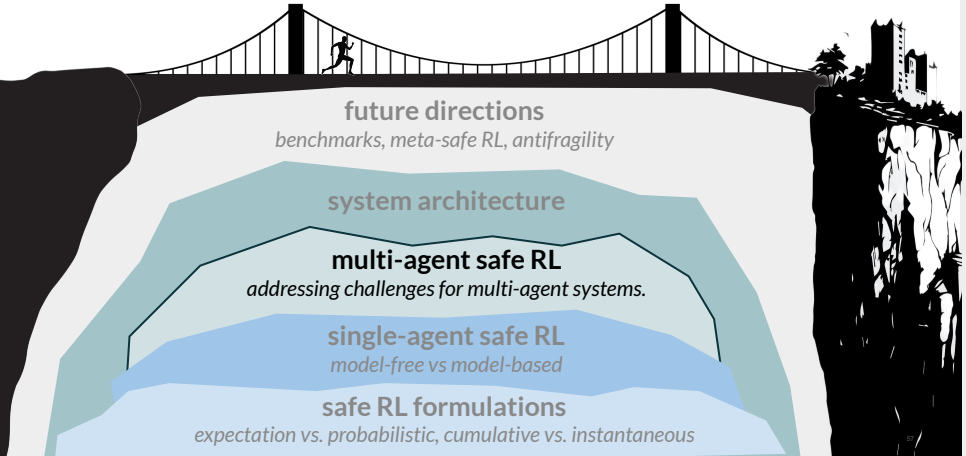
Proposed RLb



safe multi-agent RL

methodology

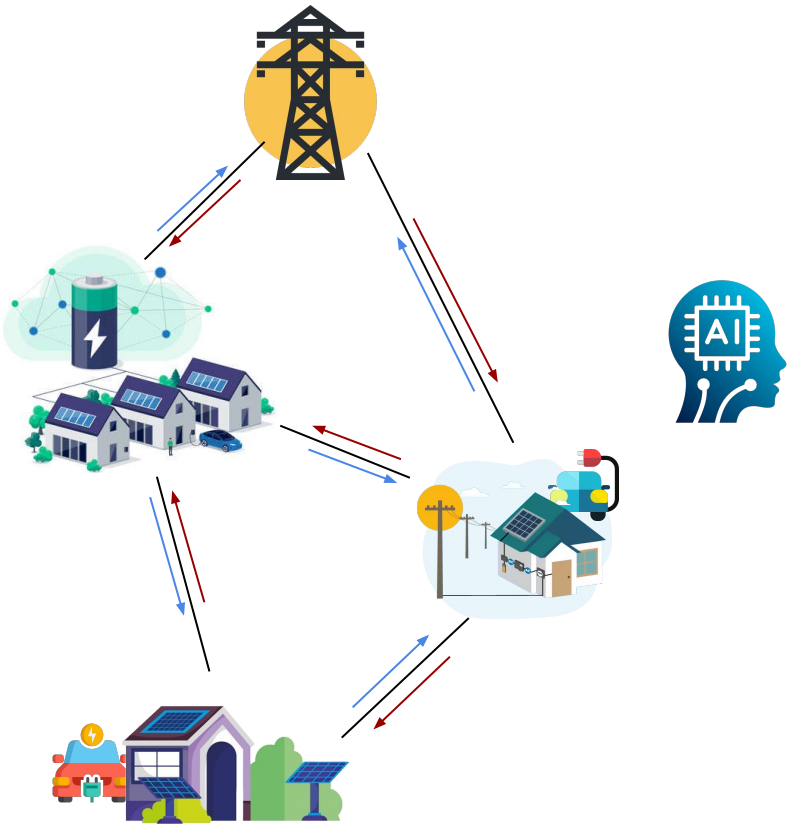
presented by Ming Jin



Key Objectives

- understand the three key challenges of MARL: perception, learning, and interaction challenges, and the approaches addresssing them
- examine the concept of safety in MARL, including global and local safety definitions
- analyze advanced techniques for safe MARL, including Lagrangian methods and control-theoretic approaches

Multi-Agent Systems



MDPs (Markov Decision Processes):
formalizing single-agent (e.g., solar panel) decisions

Goal: find the optimal policy that maximizes the **expected cumulative reward over time.**

$$\pi^* = \arg\max_\pi \mathbb{E}\left[\sum_t \gamma^t r(s_t, a_t)\right]$$

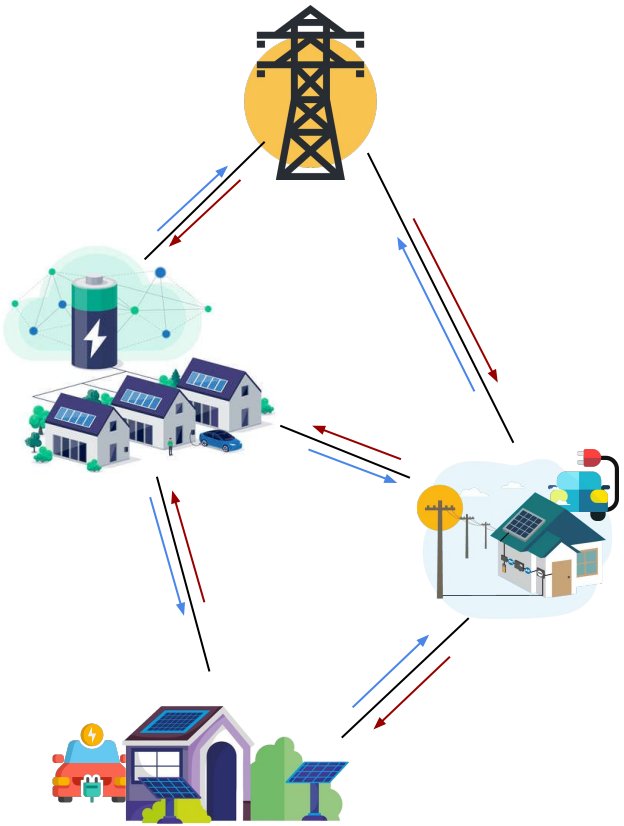
$\mathcal{S}$ **state:** the current snapshot of the system (e.g., voltage at bus)

$\mathcal{A}$ **action:** the choices the agent can make (e.g., adjusting active/reactive power injections)

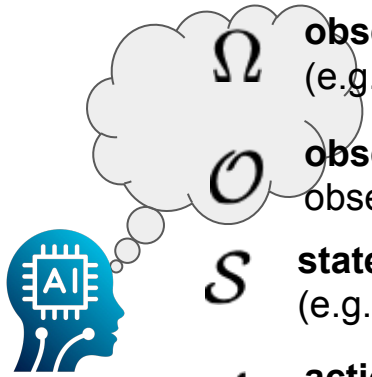
r **reward:** the feedback that guides learning (e.g., positive for regulating the voltage)

$\mathcal{P}$ **transition:** how actions change the state (e.g., AC power flows)

Multi-Agent Systems



POMDPs (Partially Observable MDPs)
single-agent with *incomplete* information about states



Ω **observation:** the set of all possible measurements (e.g., voltage, current, communication signals)

$\mathcal{O}$ **observation probability:** the probability of observation given that the system state and action

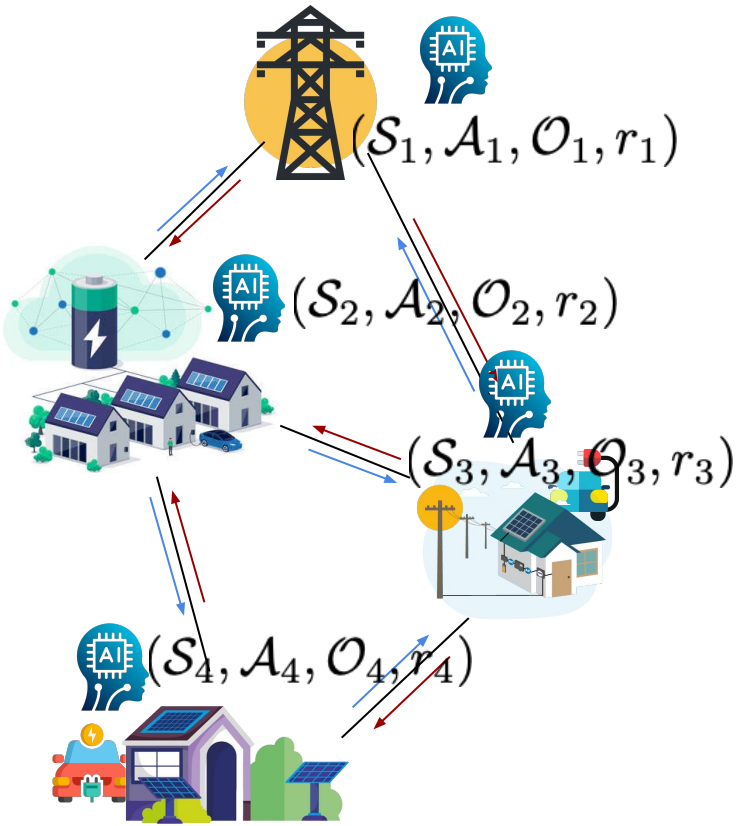
$\mathcal{S}$ **state:** the current snapshot of the system (e.g., voltage at bus)

$\mathcal{A}$ **action:** the choices the agent can make (e.g., adjusting active/reactive power injections)

r **reward:** the feedback that guides learning (e.g., positive for regulating the voltage)

$\mathcal{P}$ **transition:** how actions change the state (e.g., AC power flows)

Multi-Agent Systems



Dec-POMDPs (Decentralized POMDPs)

a framework for modeling multi-agent systems where agents have limited information with **individual rewards**

Goal: find the optimal **joint** policy that maximizes the **expected cumulative reward over time**.

$$\pi^* = \arg\max_\pi \mathbb{E}\left[\sum_t \gamma^t r(s_t, a_t)\right]$$

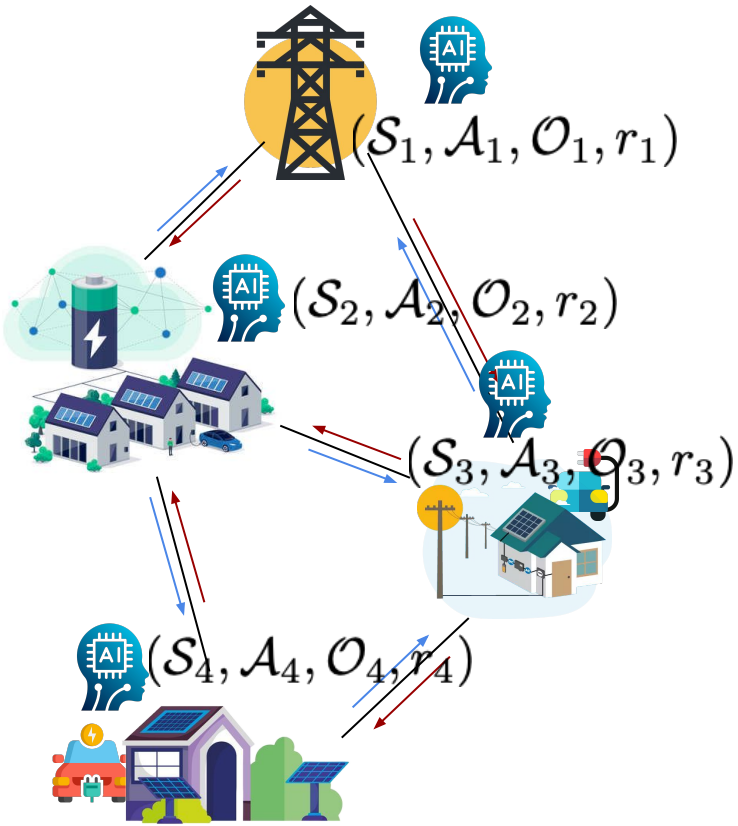
$\mathcal{K}$ **set of agents** (e.g., gid, battery, solar panel)

$$\mathcal{O} = \times_k \mathcal{O}_k$$
$$\mathcal{S} = \times_k \mathcal{S}_k$$
$$\mathcal{A} = \times_k \mathcal{A}_k$$

joint observation, state, and action spaces

$(r_1, \ldots, r_K)$ common reward for **fully cooperative** Dec-POMDP

Multi-Agent Reinforcement Learning (MARL)



Dec-POMDPs (Decentralized POMDPs)
a framework for modeling multi-agent systems where agents have limited information with **individual rewards**

Goal: find the optimal **joint** policy that maximizes the **expected cumulative reward over time**.

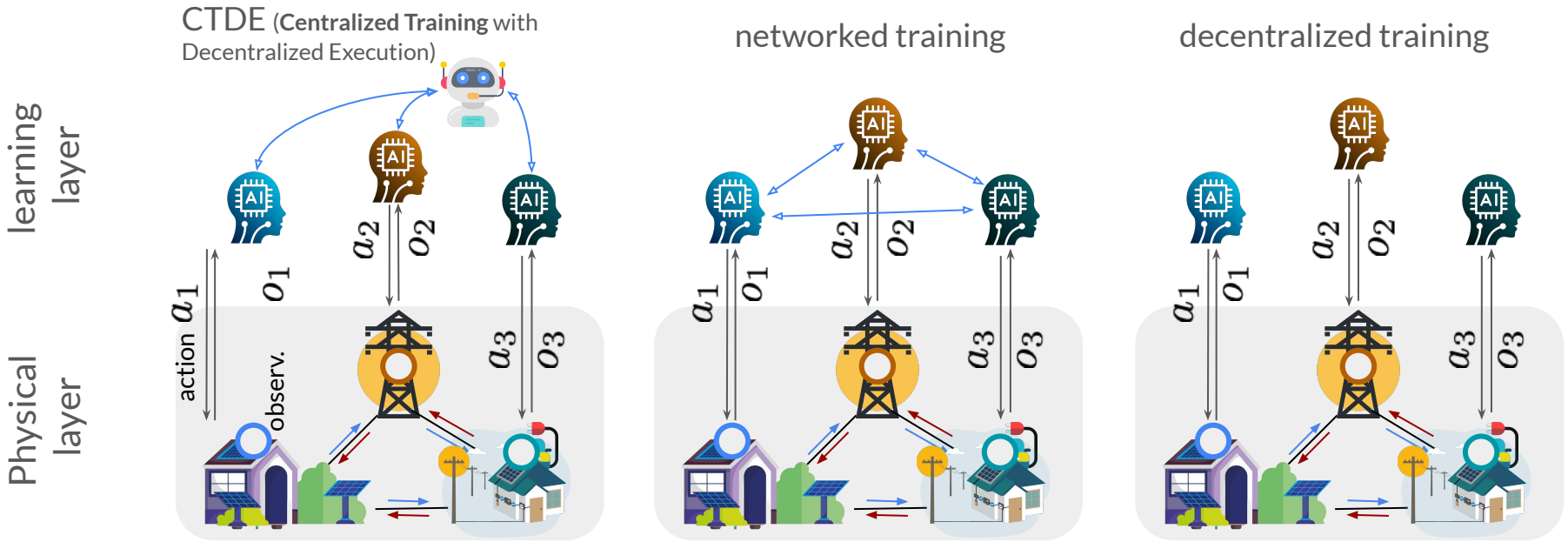
$$\pi^* = \arg\max_\pi \mathbb{E}\left[\sum_t \gamma^t r(s_t, a_t)\right]$$

MARL: multiple agents learn simultaneously in shared environment for solving Dec-POMDP

e.g., **IQL** (Independent Q-Learning), **MAAC** (Multi-Agent Actor-Critic), **MADDPG** (Multi-Agent Deep Deterministic Policy Gradient), **QMIX**

…enables these agents to learn and adapt to complex, dynamic environments

Training Regimes for MARL



	CTDE (Centralized Training with Decentralized Execution)	networked training	decentralized training
learning speed	high	medium	low
comm. require.	high	medium	low
scalability	low	medium	high

The MARL Maze: Navigating the Challenges

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***

Tackling Perception Challenge (1/3)


partial observability


non-stationarity

RNNs and memory-based approaches

maintains an internal memory of histories

integrates past information with current observations
helps capture latent state and predict future trends

e.g., PowerNet [Chen et al., 2021], Graph-MARL[Mu et al., 2023], Deep Recurrent Q-Net. [Hausknecht & Stone, 2015], Recurrent-MADDPG [Suttle et al., 2019]

PROS: relatively simple to implement, GRUs (e.g., Graph-MARL}, LSTMs (e.g., PowerNet)

CONS: can be difficult to train, esp. for long sequences

Attention mechanisms

attends to relevant parts of observations

filters out noise and prioritizes critical information

e.g., MAAC (Multi-Actor-Attention-Critic) [Iqbal & Sha, 2019], Attention-based Optimizer for Continuous Control [Jiang & Lu, 2018]

PROS: efficient for large state spaces and long-range dependencies

CONS: can be computationally expensive or over-flexible

Agent modeling and belief states

explicitly model other agents' behaviors or maintain beliefs about the true state of the environment

e.g., confederate image technology [Li et al., 2024], DRON [He et al., 2016], Bayes-Adaptive POMDP [Ross et al., 2011]

PROS: theoretically sound, effective for systems with a few agents

CONS: may not scale well, sensitive to modeling assumptions

Tackling Perception Challenge (2/3)


partial observability


non-stationarity

Multi-agent communication

allows agents to share information

can lead to emergent team strategies

e.g., PowerNet [Chen et al., 2021], Deep CommNet [Sukhbaatar et al., 2016], DIAL (Differentiable Inter-Agent Learning) [Foerster et al., 2016]

\+ PROS: can significantly improve performance in cooperative tasks

— CONS: learned protocols may not be interpretable

CTDE (Central. Training with Decentral. Execution)

leverages global information during training

agents can learn to coordinate more effectively

e.g., most MARL methods in power system literature, MADDPG and variants

\+ PROS: can stabilize training

— CONS: may not fully capture decentralized dynamics

Experience Replay Modifications

store and replay sequences of observations to help with temporal dependencies

e.g., Importance Sampling [Foerster et al., 2017], Multi-agent Fingerprints [Foerster et al., 2017]

\+ PROS: helps stabilize learning and sample efficiency, can combine with other methods

— CONS: may slow down learning in rapidly changing environments

Tackling Perception Challenge (3/3)

 partial observability  non-stationarity

	Scalability *ability to handle increasing # of agents*	**Comp. Complex.** *resource requirements for training & execution*	**Sample Efficiency** *learning from limited interactions*	**Interpretability** *ease of explaining the learned policies/behaviors*
Memory-based (RNNs)	◐		⃠	
Attention mechanisms		⃠	⃠	◐
Agent modeling/belief state	⃠		◐	◐
Communication	◐		◐	
CTDE	●	◐	●	
Experience replay mod.	◐	◐	●	

⃠ limitation ◐ good ● excellent

Key takeaways:

- most address both partial observability and non-stationarity, but no single "best" method
- consider tradeoffs based on specific problems (e.g., #agents, available data/compute)
- hybrid approaches (combine CTDE with other methods) can be effective.

Enhancing Learning Efficiency (1/3)


credit assignment


explore vs. exploit

…leveraging shared methods for synergistic improvement

1

CTDE
(Central. Training w/ Decentral. Execution)

more accurate individual contributions and coordinate exploration strategies

e.g., **central critic:** MADDPG, COMA with Exploration Curriculum [Wang et al., 2020]
central latent space: MAVEN (Multi-Agent Variational Exploration) [Mahajan et al., 2019]

2

Attention mechanism

focus on relevant information, weighing contributions and guiding exploration

e.g., MAAC use attention-based critic to implicitly assign credits

3

Communi cation

facilitate the sharing of reward information and coordinated exploration

e.g., MADDPG-M (MADDPG with Multi-Agent Coordinated Exploration) [Chu et al., 2020] uses communication specifically for coordinating exploration

Enhancing Learning Efficiency (2/3)


credit assignment


explore vs. exploit

Credit Assignment Approaches

Difference rewards/Shapley value

quantifying Individual Impact

by comparing the overall reward with and without their actions

e.g., Wonderful Life Utility [Wolpert & Tumer, 2002], Difference Rewards [Tumer & Agogino, 2007], Shapley Value based, e.g., SQDDPG [Wang et al., 2020]

PROS clear, intuitive, theoretically sound, effective for cooperative tasks

CONS not easily scalable, may struggle with interdependent tasks.

Value decomposition

decomposes the team's value function into individual agent value functions

e.g., VDN (Value Decomposition Networks) [Sunehag et al., 2018], QMIX [Rashid et al., 2018], MAAC [Lowe et al., 2017]

PROS scalable, allows for decentralized execution, flexible (QMIX)

CONS assumes additivity (VDN) or monotonicity (QMIX), may struggle with non-linear interactions

Enhancing Learning Efficiency (3/3)


credit assignment


explore vs. exploit

Exploration Approaches

Diversity-driven exploration/ parameter space exploration (PSN)

promote diverse behaviors and actions within and across agents

e.g., NoisyNet adapted for MARL (adds noise to weights) [Plappert et al., 2018], Population-Based Training with Diversity [Jaderberg et al., 2019]

\+ PROS: PSN useful for continuous action spaces, avoid premature convergence

— CONS: requires diversity metric, can be computationally demanding.

Safe exploration

explore while adhering to safety constraints

e.g., Safe MARL with Safety Constraints [Castellini et al., 2021]

\+ PROS: enables learning in high-stakes, risk-sensitive environments, provides safety guarantees during training

— CONS: can limit exploration space, may increase computational complexity due to constraint checking

Addressing Scalability Challenge (1/2)


scalability


coordination

1

Graph neural networks

model agent interactions using graphs

PROS naturally models spatial relationships, scales to different numbers of agents

e.g., DGN (Graph Convolutional RL) [Jiang et al., 2020], MAGNet (Multi-Agent Graph Network) [Malysheva et al., 2019], MAGEC [Goeckner et al., 2024]

CONS can be computationally expensive

2

Hierarchical methods

use multi-level structures for coordination

agents can learn to coordinate more effectively

PROS enables multi-level coordination

e.g., HAMA (Hierarchical Multiagent Framework) [Tang et al., 2018], FEU (Feudal Ensemble Update) [Ahuja et al., 2019]

CONS designing effective hierarchies can be challenging

3

Mean field methods

approximate interactions with many agents using average effects

PROS extremely scalable to large numbers of agents, computationally efficient

e.g., Mean Field MARL [Yang et al., 2018]

CONS may oversimplify scenarios w/ heterogeneous agents

Addressing Scalability Challenge (2/2)


scalability


coordination

1

Decomposition

value decomp. factored MDPs

Exploit problem structure to factor state and action spaces or decompose value functions

e.g., VDN, QMIX, Factored-Value MARL
[Bargiacchi et al., 2018]

2

Attention mechanism

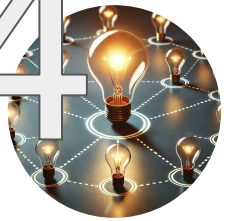
selectively focus on relevant agents or information.

e.g., MAAC (attention in critic), ATOC (attentional communication)

3

Communi cation

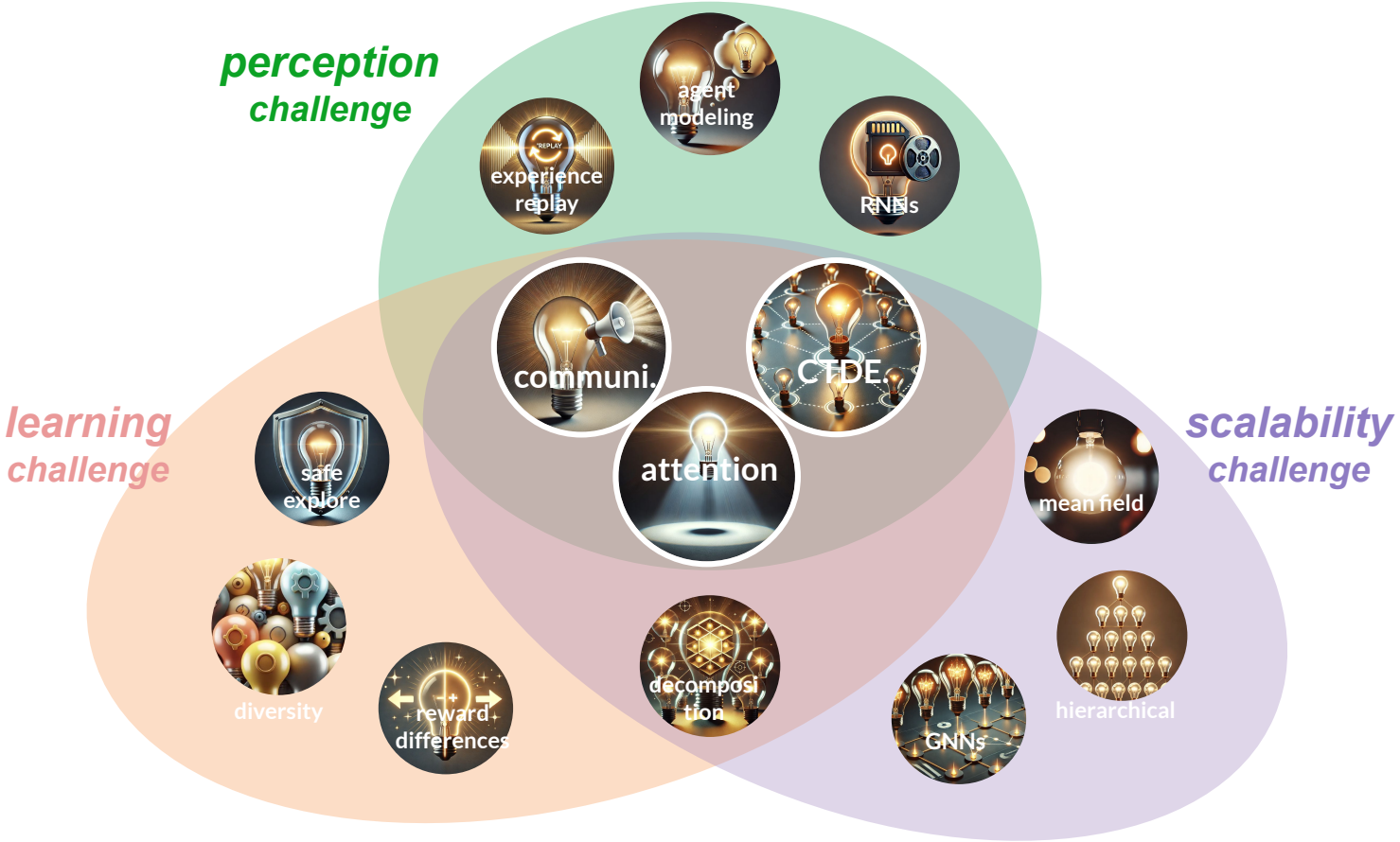
4



CTDE
(Central. Training w/ Decentral. Execution)

e.g., parameter sharing [Mu et al., 2023] [Sun and Lu, 2024]

MARL Methods: A Multi-Purpose Toolkit



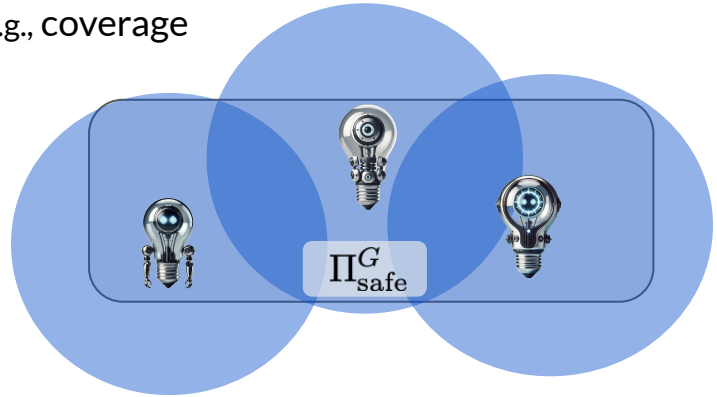
Multi-Agent Safety Definitions

Global safety

treats entire multi-agent system as one entity

joint policy constraint: $\pi = (\pi_1, ..., \pi_K) \in \Pi^G_{\text{safe}}$

E.g., coverage

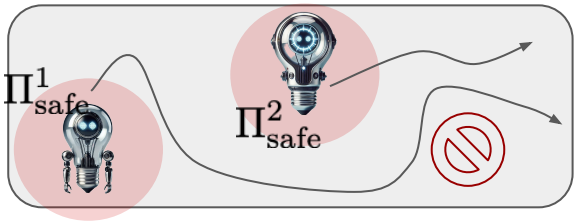


Local safety

each agent has its own safety constraint

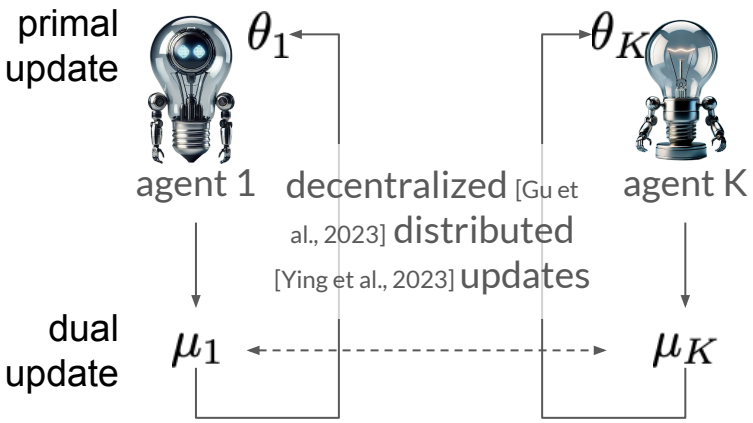
individual constraints: $\pi_k \in \Pi^k_{\text{safe}}, \quad \forall k$

E.g., collision avoidance



♠ **smart grid:** global (system stability) vs local (voltage limits)

Lagrangian Methods for Safe MARL



safe MARL formulation with **general utilities**

$$\max_{\theta:=\{\theta_k\}_{k\in\mathcal{K}}} F(\theta) = \sum_k f_k(\lambda^{\theta_k})$$

$$\text{s.t. } G_k(\theta) = g_k(\lambda^{\theta_k}) \le 0 \quad \forall k \in \mathcal{K}$$

local occupancy measure of agent k

joint policy parameters

local objective/constraint

*general utilities f_k and g_k can be arbitrary (potentially nonlinear) functions for imitation learning, exploration.

*standard cumulative rewards/costs:

$f_k(\lambda^{\pi_k}) = \langle r_k, \lambda^{\pi_k} \rangle$ (where r_k is the local reward vector)

Lagrangian formulation w/ primal-dual updates

$$\max_{\theta} \min_{\mu \ge 0} L(\theta, \mu) = F(\theta) + \sum_k \mu_k G_k(\theta)$$

- primal update: each agent updates local policy θ_k
- dual updates: centralized [Gu et al., 2023] or decentralized (using shadow rewards and m-hop policy truncation) [Ying et al., 2023]

Distributed Optimization for Safe MARL

- use distributed optimization for constrained policy optimization
- incorporate safety constraints as coupling constraints across agents

CBF-induced central filter

$$\min_{u_1,\dots,u_K} \sum_{k=1}^{K} \frac{1}{2}\|u_k - u_{\text{nom},k}\|^2$$

$$\text{s.t.} \sum_{k=1}^{K} A_k u_k = c$$

♠ introduce auxiliary y_k variable for each agent k
♠ local exchange of auxiliary variables

auxiliary variable decomposition

for each agent k:

$$\min_{u_k} \|u_k - u_{\text{nom},k}\|^2$$

$$\text{s.t. } A_k u_k + \sum_{j \in N_k} (y_k - y_j) = c$$

? distributed decomposition, e.g., ADMM (Lagrangian-based)

key challenge: maintain safety during **every iteration** (violation-free/feasible algorithms)

safety at every iteration: sum of local constraints across all agents equals the global constraint:

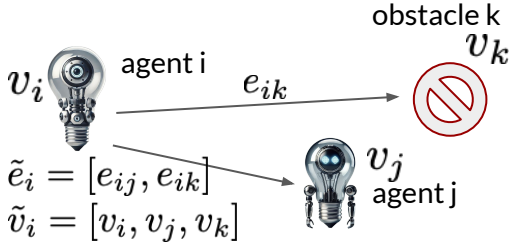
$$\sum_{k=1}^{K} \sum_{j \in N_k} (y_k - y_j) = 0$$

Tan and Dimarogonas (2022), Bai et al. (2024)

future directions: integration with deep RL, robust formulations for uncertain dynamics

Graph CBFs for MAS

- learning-based approach using Graph Neural Networks
- handles general nonlinear dynamics: $\dot{x}_k = f_k(x_k, u_k)$
- key advantages:
  - scalable to large numbers of agents
  - handles variable number of neighbors
  - distinguishes between agent types
  - generalizes to unseen scenarios



Graph CBF conditions

$h(\tilde{e}_i, \tilde{v}_i) > 0$ for all safe states

$h(\tilde{e}_i, \tilde{v}_i) < 0$ for all unsafe states

$h(\tilde{e}_i, \tilde{v}_i) \geq -\alpha(h(\tilde{e}_i, \tilde{v}_i))$ for all safe states

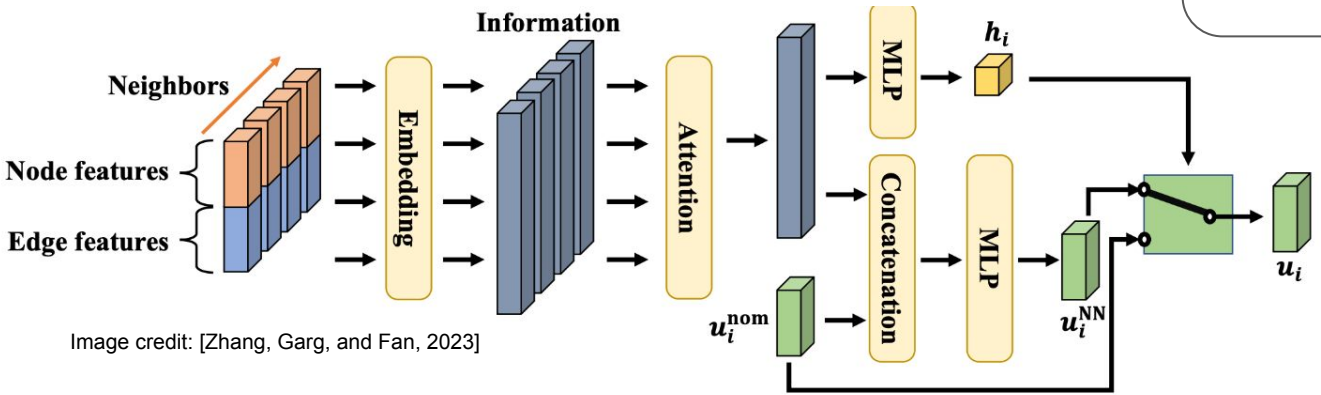


Image credit: [Zhang, Garg, and Fan, 2023]

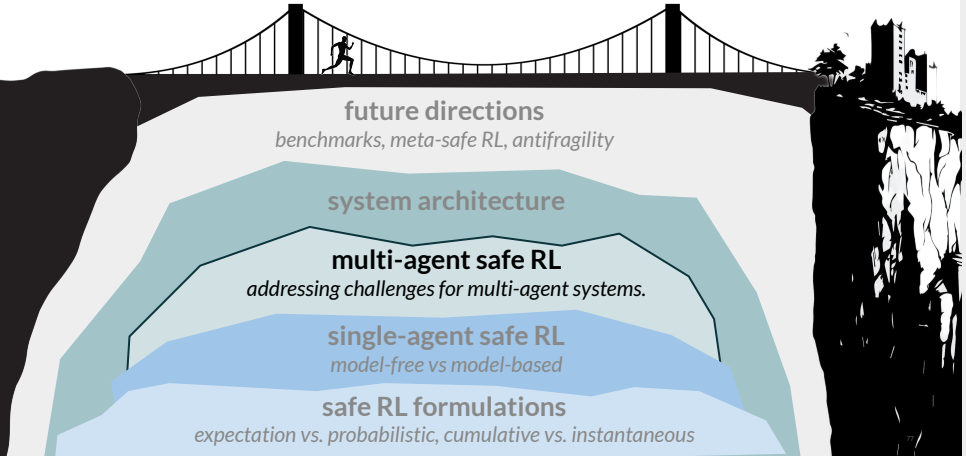
Trained with only 16 agents, GCBF can still maintain more than 50% success rates for >1000 agents when other methods completely fail.

See also: [Tolstaya et al., 2021], [Li et al., 2020], [Yu et al., 2023], [Blumenkamp et al., 2022], [Jia et al., 2022], [Ji et al., 2021]

safe multi-agent RL

case studies

presented by Vanshaj Khattar



Key Objectives

- understand the three key challenges of MARL: perception, learning, and interaction challenges, and the approaches addresssing them
- examine the concept of safety in MARL, including global and local safety definitions
- analyze advanced techniques for safe MARL, including Lagrangian methods and control-theoretic approaches

Application: EV charging in microgrids

Goal: Reach desired SOC with minimum charging expenses

Constraints: Charging discharging capacity, load balancing constraints

A case study on using **deep-RL with safety filter for** EV charging problem

- time horizon: $t \in \mathcal{T} = \{1, 2, \ldots, T\}$
- charging piles: $n \in \mathcal{N}_c = \{1, 2, \ldots, N\}$
- remaining demand: $d_{n,t}^{\text{res}}$
- residual parking time: $h_{n,t}^{\text{res}}$
- decision interval: Δt
- electricity generated by PV: $E_{\text{PV},t}$
- residential load demand: $E_{\text{load},t}$
- price charged by power grid: PG_t
- price charged by the microgrid: PS_t
- price charge by EVs to microgrids: PB_t
- charging and discharging actions: $a_{\text{BES},t}$
- electricity purchased/sold from grid: $a_{G,t}$
- n-th pile charging/discharging action: $a_{n,t}$

Zhang, Shulei, et al.

Application: EV charging

CMDP formulation:

State: $\{t, E_{PV,t}, E_{\text{load},t}, E_{\text{BES},t}, d^{\text{res}}_{1,t}, h^{\text{res}}_{1,t}, \ldots, d^{\text{res}}_{n,t}, h^{\text{res}}_{n,t} \ldots, d^{\text{res}}_{N,t}, h^{\text{res}}_{N,t}, PG_t, PS_t, PB_t\}$

Actions: $\{a_{\text{BES},t}, a_{G,t}, a_{1,t}, \ldots, a_{n,t}, \ldots, a_{N,t}\}$

Constraints: Charging/discharging limits: $\sum_{n=1}^{N} a_{n,t} z^{c}_{n,t} \leq e^{c}_{up}$ $\sum_{n=1}^{N} a_{n,t} z^{d}_{n,t} \leq e^{d}_{up}$

Each EV demand demand should be met: $\sum_{t^a_n=t}^{t^a_n+h^{\text{res}}_{n,t}} a_{n,t^a_n} z^{c}_{n,t^a_n} \geq d^{\text{res}}_{n,t} + \sum_{t^a_n=t}^{t^a_n+h^{\text{res}}_{n,t}} a_{n,t^a_n} z^{d}_{n,t^a_n}$

BES hard constraints: $-e^{\text{BES}}_{up} \leq a_{\text{BES},t} \leq e^{\text{BES}}_{up}$ + $0 \leq E_{\text{BES},t} + a_{\text{BES},t} \leq E^{up}_{\text{BES}}$

Application: EV charging

CMDP formulation:

Reward:

$$\underbrace{\mathrm{Pr}_t(A_t) = PG_t E_{\mathrm{load},t} + PS_t \sum_{n=1}^{N}(a_{n,t} z^c_{n,t}) - PG_t a_{G,t} - PB_t \sum_{n=1}^{N}(a_{n,t} z^d_{n,t})}_{\text{Profit at time step t}} \Rightarrow r_t(A_t, S_t) = \begin{cases} 0, & \text{if } t < T \\ \sum_{t=1}^{T} \mathrm{Pr}_t, & \text{if } t = T \end{cases}$$

Safe RL approach: → Constrained SAC (CSAC)

$$Q_\pi(s,a) = \mathbb{E}_{(s_t,a_t)\sim\rho\pi}\left[\sum_{t=0}^{T}\gamma^t r(s_t,a_t) + \alpha\sum_{t=1}^{T}\gamma^t \mathcal{H}(\pi(\cdot|s_t))|s_0 = s, a_0 = a\right]$$

$$V_\pi(s) = \mathbb{E}_{(s_t,a_t)\sim\rho\pi}\left[\sum_{t=0}^{T}\gamma^t r(s_t,a_t) + \alpha\gamma^t \mathcal{H}(\pi(\cdot|s_t)) \mid s_0 = s\right]$$

Penalty function design

$$c(s_t,a_t) = \underbrace{\max(0, -E_{\mathrm{BES},t} - a_{\mathrm{BES},t}) + \max(0, a_{\mathrm{BES},t} - E^{up}_{\mathrm{BES}} + E_{\mathrm{BES},t})}_{term1} + \underbrace{\omega_{c1}\max(0, \beta_t - e_t) + \max(0, e_t - \zeta_t)}_{term2} + \underbrace{\omega_{c2} E_{\mathrm{BES},T}}_{term3}$$

Application: EV charging

Constrained formulation

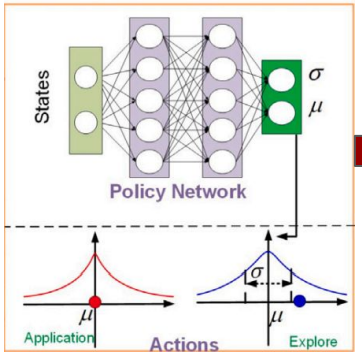
$$\pi_t^* = \arg\max_{\pi_t \in \Pi} V_\pi(s_t)$$
$$= \arg\max_{\pi_t \in \Pi} \mathbb{E}_{a_t \sim \pi} \left[Q_\pi(s_t, a_t) + \alpha \mathcal{H}(\pi(a_t|s_t)) \right]$$
$$\text{s.t. } V_\pi^c(s_t) \leq d$$

Unconstrained formulation

$$\pi_t^* = \mathcal{L}(\pi_t, \lambda) = \arg\min_{\lambda_0} \max_{\pi_t \in \Pi} \left[V_\pi(s_t) - \lambda (V_\pi^c(s_t) - d) \right]$$
$$= \arg\min_{\lambda_0} \max_{\pi_t \in \Pi} \mathbb{E}_{a_t \sim \pi} \left[\begin{matrix} Q_\pi(s_t, a_t) + \alpha \mathcal{H}(\pi(a_t|s_t)) \\ - \lambda (Q_\pi^c(s_t) - d) \end{matrix} \right]$$

Guaranteed constraint satisfaction through safety filter:

- Rule-based safety filter is introduced



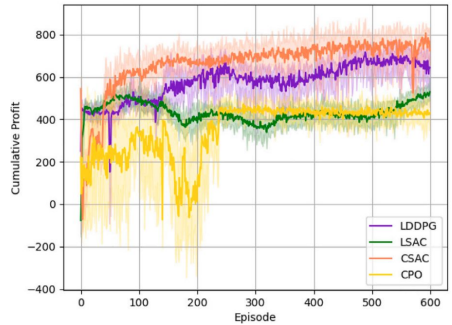
if $e_t < \beta_t$ **then**
  $e_t = \beta_t$
else if $e_t > \zeta_t$ **then**
  $e_t = \zeta_t$
else
  $e_t = e_t$
end if
if $a_{\text{BES},t} + E_{\text{BES},t} < 0$ **then**
  $a_{\text{BES},t} = -E_{\text{BES},t}$
else if $a_{\text{BES},t} + E_{\text{BES},t} > E_{\text{BES}}^{up}$ **then**
  $a_{\text{BES},t} = E_{\text{BES}}^{up} - E_{\text{BES},t}$
else
  $a_{\text{BES},t} = a_{\text{BES},t}$
end if

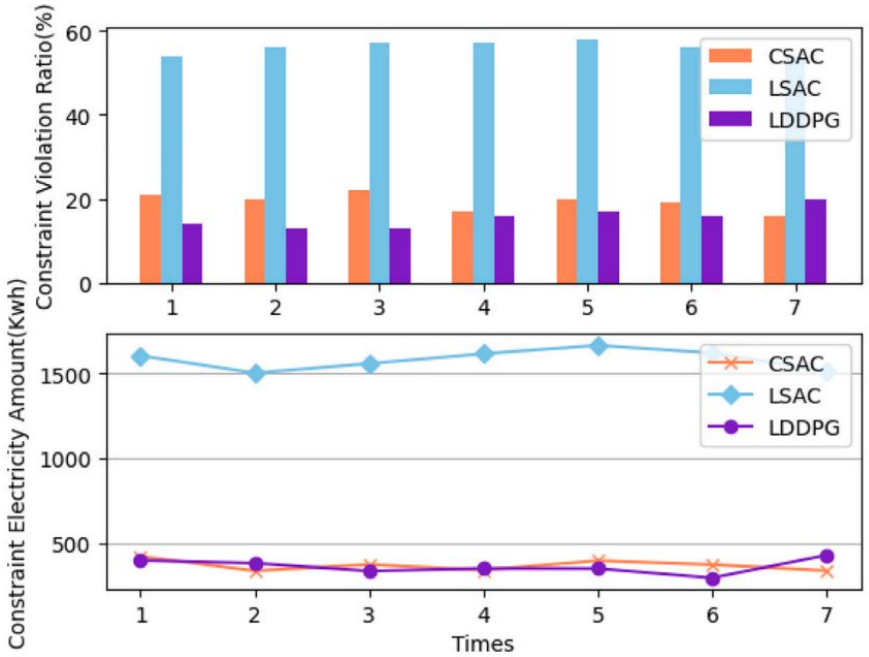
Safe actions

Application: EV charging

Case study

- Residential microgrid equipped with a PV, BES, residential loads, and a large charging station
- Three different types of EVs are considered





Application: Distributed load restoration

Goal: Restore as many critical loads as possible using local DERs under extreme events

Power flow constraints, ESS constraints, MT constraints

We present a case study on using **model-free safe RL** method based on **Constrained Policy Optimization (CPO)** for distributed load restoration.

- time horizon: $t \in \mathcal{T} = \{1, 2, \ldots, T\}$
- complex power injection: $s_{i,t}$
- squared magnitude of voltage phasor: $v_{i,t}$
- load buses: $\mathcal{N}^{\mathrm{L}}$
- buses connected to turbine: $\mathcal{N}^{\mathrm{MT}}$
- Convex cost functions: C^{L}, C^{MT}

Load restoration objective

$$J_{\mathcal{T}} \triangleq \sum_{t \in \mathcal{T}} \left(\sum_{i \in \mathcal{N}^{\mathrm{L}}} C^{\mathrm{L}}_{i,t}(\Re(s_{i,t})) + \sum_{i \in \mathcal{N}^{\mathrm{MT}}} C^{\mathrm{MT}}_{i,t}(\Re(s_{i,t})) \right)$$

under power flow constraints, MT constraints, ESS constraints, RES and load constraints

restoration horizon

Number of loads

Number of turbines

Bose, Shourya et.al. (2023)

Application: Distributed load restoration

CMDP formulation of the LR problem

State:

$$\left\{ \{\mathcal{S}_{i,t-1}\}_{i\in\mathcal{N}^{\text{ESS}}}, \{\zeta_{i,t-1}\}_{i\in\mathcal{N}^{\text{MT}}}, \{\rho_{i,t-1}\}_{i\in\mathcal{N}^{\text{L}}}, \left[\left\{ \hat{P}^{\text{L}}_{i,t'}, \hat{Q}^{\text{L}}_{i,t'} \right\}_{i\in\mathcal{N}^{\text{L}}}, \left\{ \hat{P}^{\text{RES}}_{i,t'} \right\}_{i\in\mathcal{N}^{\text{RES}}} \right]_{t'=t}^{t+H-1} \right\}$$

Action:

$$\left\{ \left[\{s_{i,t'}\}_{i\in\mathcal{N}^{\text{RES}}}, \{s_{i,t'}\}_{i\in\mathcal{N}^{\text{MT}}}, \{\rho_{i,t'}\}_{i\in\mathcal{N}^{\text{L}}} \right. \right. \left. \left. \left\{ P^{\text{ch}}_{i,t'}, P^{\text{dis}}_{i,t'}, \Im(s_{i,t'}) \right\}_{i\in\mathcal{N}^{\text{ESS}}} \right]_{t'=t}^{t+H-1} \right\}.$$

Reward:

$$R(\mathbf{x}_t, \mathbf{u}_t) = \sum_{t'=t}^{t+H} \gamma^{(t-t')} \left[\sum_{i\in\mathcal{N}^{\text{L}}} C^{\text{L}}_{i,t'}(\Re(s_{i,t'})) + \sum_{i\in\mathcal{N}^{\text{MT}}} C^{\text{MT}}_{i,t'}(\Re(s_{i,t'})) \right].$$

Reward and constraint function

$$\mathcal{J}^{R}(\pi_{\boldsymbol{\theta}}, \mathbf{x}_t) \triangleq \mathbb{E}_{\mathbf{u}_t\sim\pi_{\boldsymbol{\theta}}} \left[R(\mathbf{x}_t, \mathbf{u}_t) \,|\, \mathbf{x}_t \right]$$

$$\mathcal{J}^{\mathbf{C}}(\pi_{\boldsymbol{\theta}}, \mathbf{x}_t) \triangleq \mathbb{E}_{\mathbf{u}_t\sim\pi_{\boldsymbol{\theta}}} \left[\mathbf{C}(\mathbf{x}_t, \mathbf{u}_t) \,|\, \mathbf{x}_t \right]$$

Power flow constraints,
ESS constraints, MT constraints

Application: Distributed load restoration

Constrained Policy Optimization (CPO) for the CLR problem

$$\boldsymbol{\theta}_{t+1} = \arg\max_{\boldsymbol{\theta}} \mathcal{J}^R(\pi_{\boldsymbol{\theta}}, \mathbf{x}_t)$$
$$\text{s.t.} \quad \mathcal{J}^{\mathbf{C}}(\pi_{\boldsymbol{\theta}}, \mathbf{x}_t) \preceq \mathbf{0}$$
$$D_{\text{KL}}\left(\pi_{\boldsymbol{\theta}}(\,.\,|\mathbf{x}_t) \| \pi_{\boldsymbol{\theta}_t}(\,.\,|\mathbf{x}_t)\right) \leq \delta$$

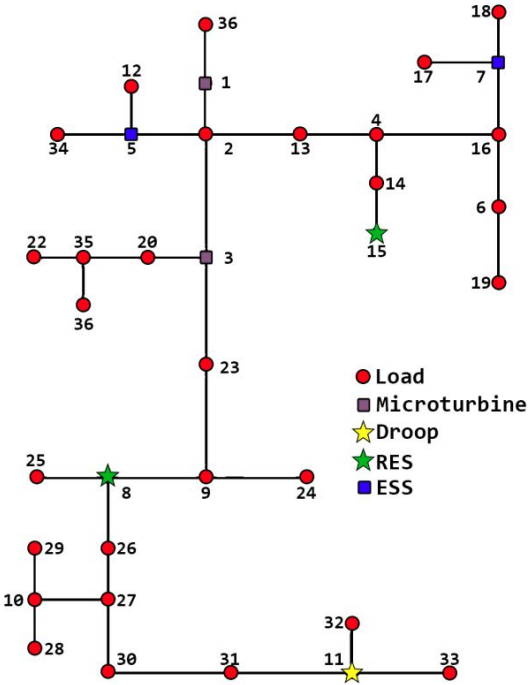
QCQP approx.

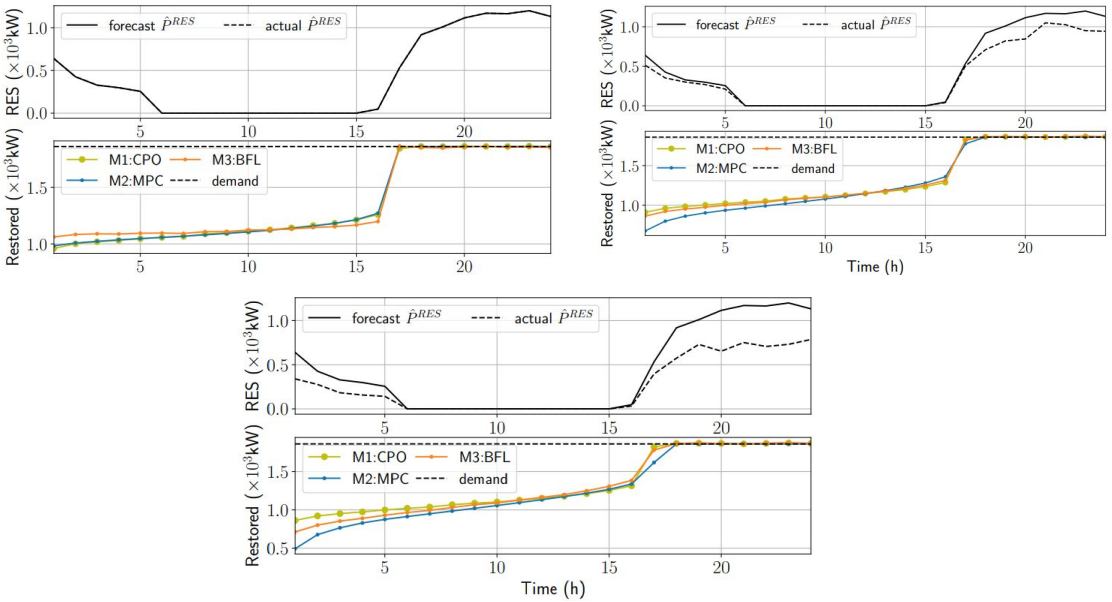
$$\boldsymbol{\theta}_{t+1} = \arg\max_{\boldsymbol{\theta}} \mathbf{a}_t^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_t)$$
$$\text{s.t.} \quad \mathbf{B}_t^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_t) + \mathbf{c}_t \preceq \mathbf{0}$$
$$(\boldsymbol{\theta} - \boldsymbol{\theta}_t)^\top \mathbf{F}_t (\boldsymbol{\theta} - \boldsymbol{\theta}_t) \leq \delta$$

$$\mathbf{a}_t = \left(\mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d^{\text{id}})} \left[\frac{\partial R_t}{\partial \mathbf{u}_t} \left((\boldsymbol{\epsilon}^\top \otimes \mathbf{I}_d^{\text{id}}) \frac{\partial \mathbf{v}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \right) \middle| \mathbf{x}_t \right] \right)_{\boldsymbol{\theta} = \boldsymbol{\theta}_t}$$
$$\mathbf{B}_t = \left(\mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d^{\text{id}})} \left[\frac{\partial \mathbf{C}_t}{\partial \mathbf{u}_t} \left((\boldsymbol{\epsilon}^\top \otimes \mathbf{I}_d^{\text{id}}) \frac{\partial \mathbf{v}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \right) \middle| \mathbf{x}_t \right] \right)_{\boldsymbol{\theta} = \boldsymbol{\theta}_t}$$
$$\mathbf{c}_t = J^{\mathbf{C}}(\pi_{\boldsymbol{\theta}_t}, \mathbf{x}_t)$$
$$\mathbf{F}_t(i,j) = \left[\frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}}^\top}{\partial \boldsymbol{\theta}(i)} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}(j)} + \frac{1}{2} \text{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}(i)} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}(j)} \right) \right]_{\boldsymbol{\theta} = \boldsymbol{\theta}_t}$$

Application: Distributed load restoration

Case study on an 36-bus microgrid

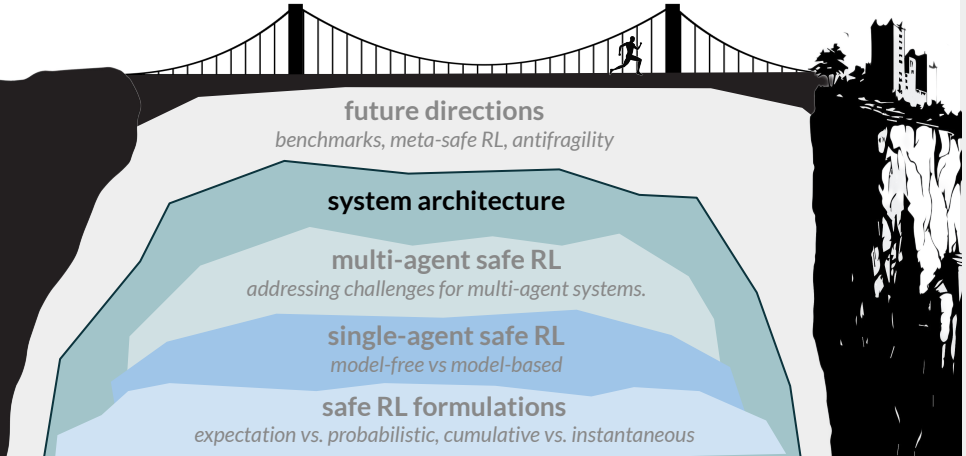




system architecture

methodology

presented by Ming Jin



Key Objectives

- understand the extreme safety requirements in critical systems and the unique challenges they pose for RL
- explore Run Time Assurance (RTA) architectures for ensuring safety in complex, unverified systems

Extreme Safety Requirements in Power Systems

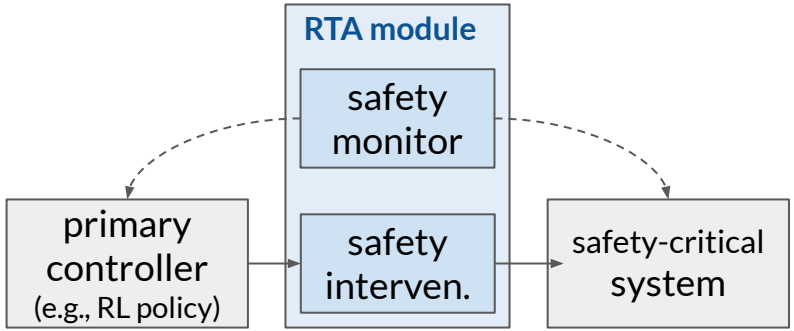
challenges specific to RL:

- exploration-exploitation dilemma
- distributional shift
- need for robust generalization

Run Time Assurance (RTA) Approaches to Safety

Key ideas:

- online **monitoring** and **intervention**
- safety filter between controller and system
- enables use of complex, unverified controllers (e.g., RL policies)



Key properties:

innocuity
no unintended negative impacts

viability
maintains future safe options

nuisance free
min. unnecessary interventions

Key industry applications: F-16 Automatic Ground Collision Avoidance System (Auto GCAS) [Swihart, et al., 2011], Mobileye's Responsibility-Sensitive Safety (RSS) model for autonomous vehicles [Shalev-Shwartz, et al., 2017]

see also [Table 1, Hobbs et al., 2023] for a survey on RTA in industry.

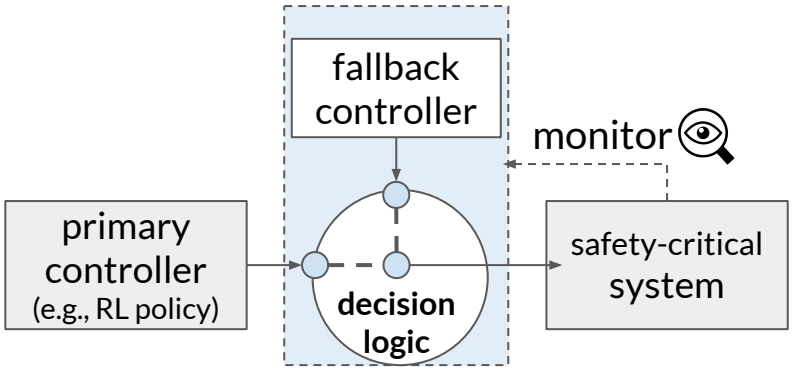
RTA Architectures for Safe Learning Systems

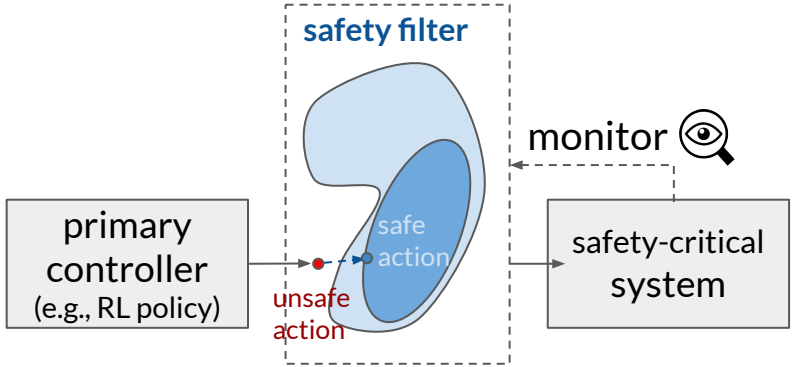
Simplex Architecture:

- **Switching** between learner and verified policy
- **Separation** of performance and safety
- Safe exploration through fallback mechanism

Safety Filtering:

- **Continuous** modification of control inputs
- **Integration** of safety into learning process
- Minimally invasive interventions



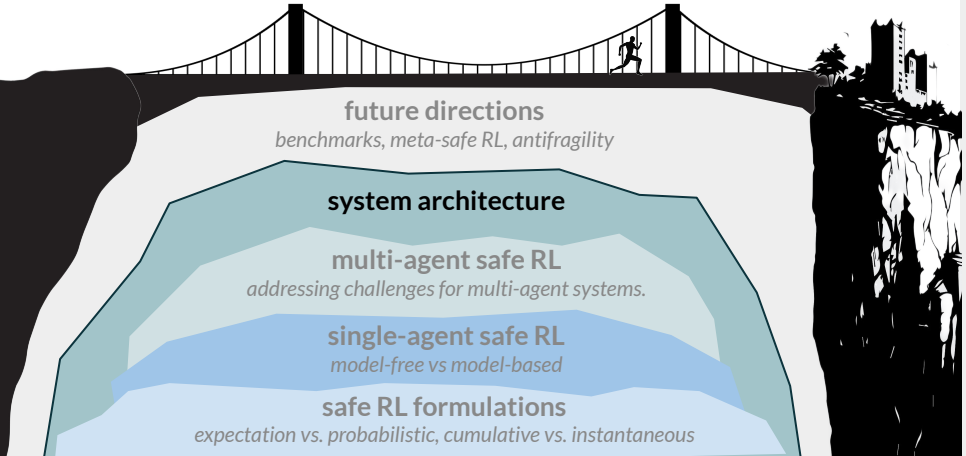


see also [Table 1, Hobbs et al., 2023] for a survey on RTA in industry.

system architecture

case studies

presented by Vanshaj Khattar



Key Objectives

- understand the extreme safety requirements in critical systems and the unique challenges they pose for RL
- explore Run Time Assurance (RTA) architectures for ensuring safety in complex, unverified systems

Application: Volt-Var control (VVC)

Goal: Maintain acceptable voltage at all points along the power network under all loading conditions

Failure can lead to **rapid voltage fluctuations and reversed power flow**

A case study for **model-based safe RL** which uses **QP-based neural network layer** to ensure safe actions.

- nodal voltage magnitude: V_t^i
- real power injection: p_t^i
- reactive power injection: q_t^i
- capacitor status time t: x_t^{cap}
- tap position time t: x_t^{reg}
- cost coefficient line loss and switching: C^l, C^s
- all device position time t: $\boldsymbol{x}_t$

VVC formulation

$$\min_{\boldsymbol{x}_t, t\in\mathcal{T}} \quad \sum_{t\in\mathcal{T}} C^l p_t^l + C^s |\boldsymbol{x}_{t-1} - \boldsymbol{x}_t|$$

$$\text{s.t.} \quad \underline{V} \le V_t^i \le \bar{V} \quad \forall t \in \mathcal{T}, i = 2, \dots, N$$

$$f_{eq}(p_t^l, q_t^l, V_t^i, x_t^{\text{cap}}, x_t^{\text{reg}}) = 0$$

N: total nodes

Gao, Yuanqi et.al.

Application: Volt-Var control (VVC)

MDP formulation of VVC

Action: $A_t = \boldsymbol{x}_t^c = [x^{c,1}, x^{c,2}, \dots, x^{c,K}]$

State: $S_t = [\boldsymbol{p}_t, \boldsymbol{q}_t, \boldsymbol{x}_{t-1}^c, t]$

Reward: $R_{t+1} = -C^l p_t^l - \sum_k C^s |[x_{t-1}^{c,k}| - [x_t^{c,k} \| - C^v C_{t+1}$

$C_{t+1} = \sum_{i=1,\dots,N} |V_t^i - 1.0|$

Model-Augmented Actor-Critic (MAAC)

- Parametric model of the VVC environment: $p_\theta(s'|s,a)$ and $r_\theta(s,a)$ where $\theta = [\theta_p, \theta_r]$
- Value function and policy parametrization: $q_\varphi(s,a)$ and $\pi_\psi^f(a|s)$

Safe exploration by embedding quadratic programming layers

$|V_t^i(s,a) - 1| < \Delta V \ \forall i \ \Leftrightarrow \ \max |V_t^i(s,a) - 1| < \Delta V$

Actor network output

$u_\psi = \pi_\psi(s) + \xi \rightarrow a_\psi = f(u_\psi)$

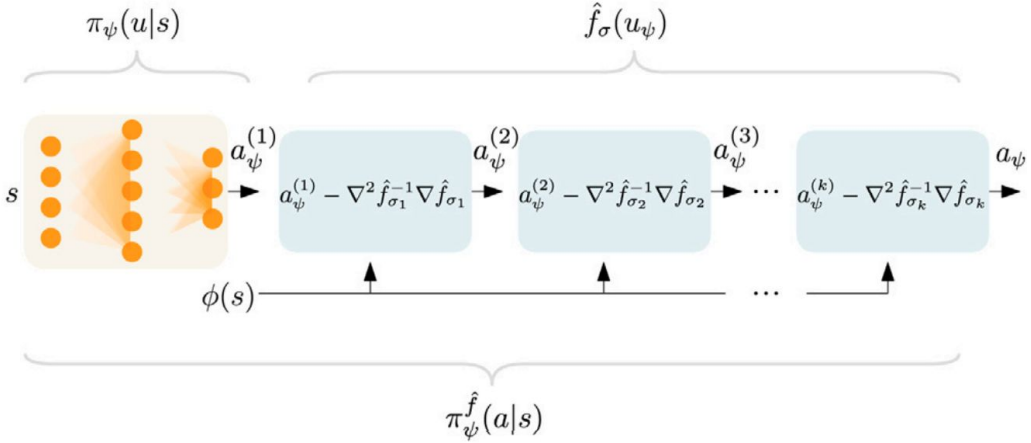
Clavera, Ignasi et.al.

Application: Volt-Var control (VVC)

$$f(u_\psi) = \underset{a}{\text{argmin}} \quad \|a - u_\psi\|^2$$
$$\text{s.t.} \quad \max_i |V^i(s,a) - 1| < \Delta V$$
$$-1 \preceq a \preceq 1$$

Estimate voltage constraint function

$$\max_i |V^i(s,a) - 1| \approx \phi_0(s) + \sum_{k=1}^{K} \phi_k(s) a_k$$



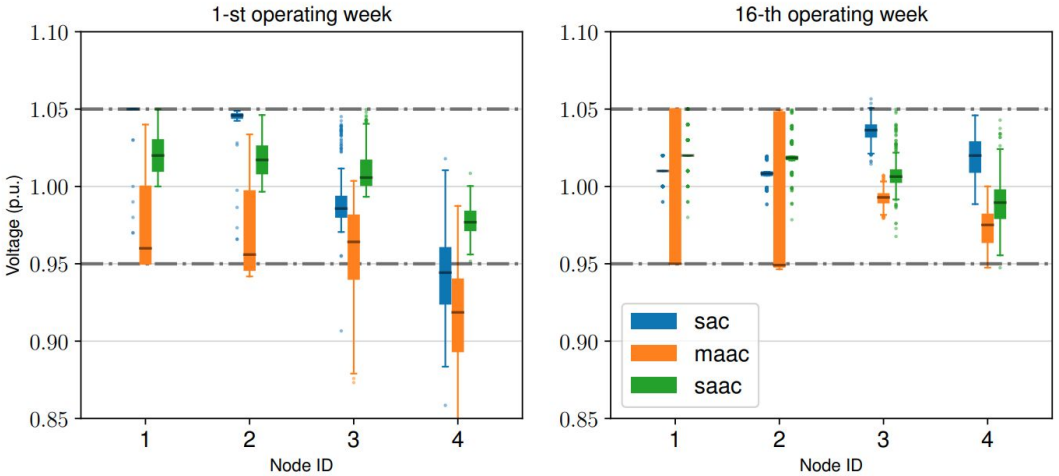
$$\hat{f}_\sigma(u_\psi) = \underset{a}{\text{argmin}} \, ||a - u_\psi||^2 + \sigma \mathbf{h}(\phi(s)^\intercal [1; a] - \Delta V) + \sigma \mathbf{h}(a - 1) + \sigma \mathbf{h}(-a - 1)$$

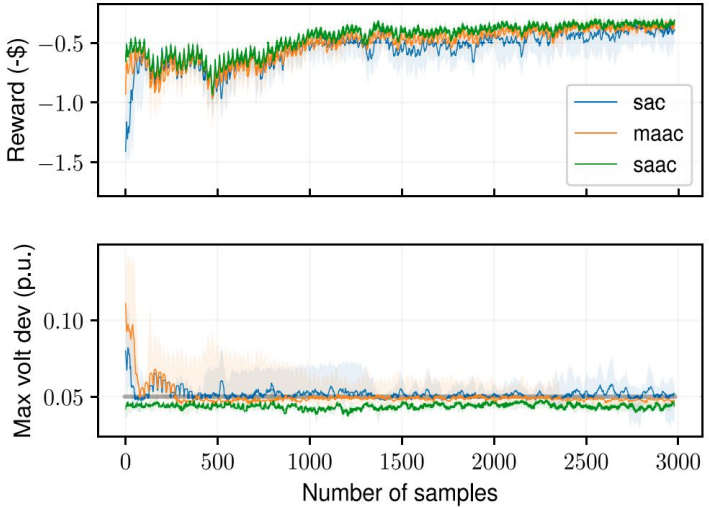
$$a_\psi^{(\nu+1)} = a_\psi^{(\nu)} - \left[\nabla^2 \hat{f}_{\sigma_\nu}(a_\psi^{(\nu)})\right]^{-1} \nabla \hat{f}_{\sigma_\nu}(a_\psi^{(\nu)})$$

Application: Volt-Var control (VVC)

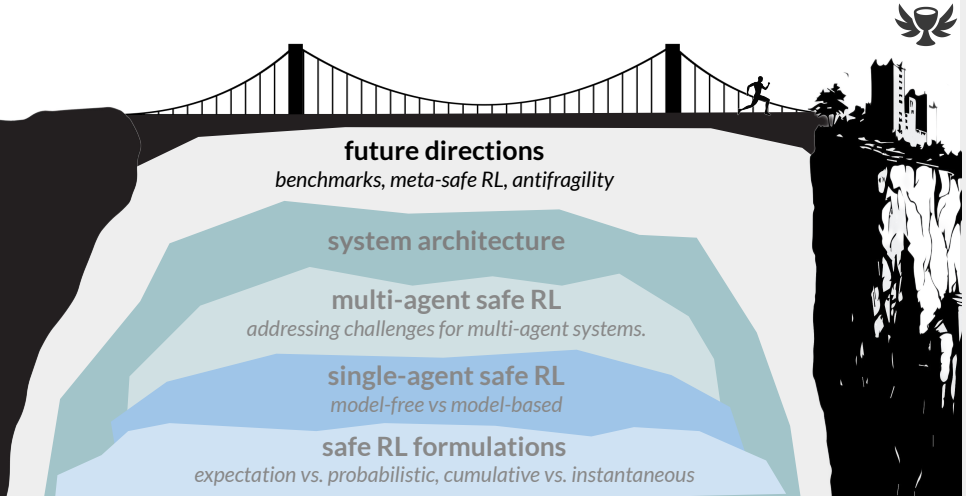
Case study on an IEEE-4 bus system

- 11 tap positions with turns ratios from 0.95 to 1.05.
- OLTCs placed at branch (2, 3) for the 4-bus feeder
- The London smart meter dataset is used to represent the nodal power injection patterns.





challenges & future directions



future research directions

- safe RL benchmarks for power systems
- from constraint satisfaction to performance optimization in open-world environments
- advancing adaptive safety in dynamic scenarios
- antifragility in AI: developing systems that improve through exposure to challenges and uncertainties

Existing benchmarks and extension to safety

SustainGym

- RL Environments for Sustainable Energy Systems
- **Key features:** Shifts in demand and environment parameters

PowerGridworld

- customizable framework for creating power-systems-focused, multi-agent Gym environments
- **Key features:** plug and play modularity, able to integrate power flow solutions

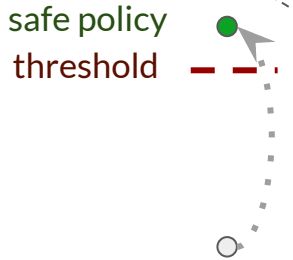
CityLearn

- RL framework for demand response and urban energy management
- **Key features:** allows for customization of the reward function, central and multi-agent mode, allows to customize the power vs. SoC,

Extension to safe-RL: safety metrics such as balancing different safety metrics. Exploration vs safety, generalization, benchmarks to test model-based safe RL methods.

Yeh, Christopher, et al., Biagioni, David, et al., Vazquez-Canteli, Jose R., et al.

Reimagining Safety in RL

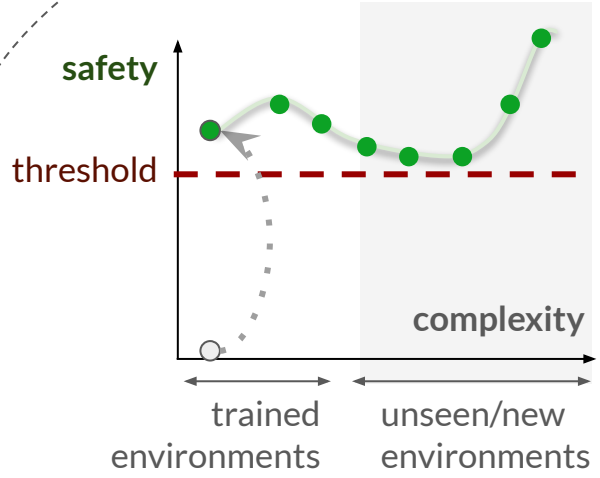


safety as constraint
in trained
environments

nonstationarity
adapting to constantly
changing environments

unknown unknowns
handling unforeseen
events

beyond robustness
thriving in adverse conditions
and safe in black swan events

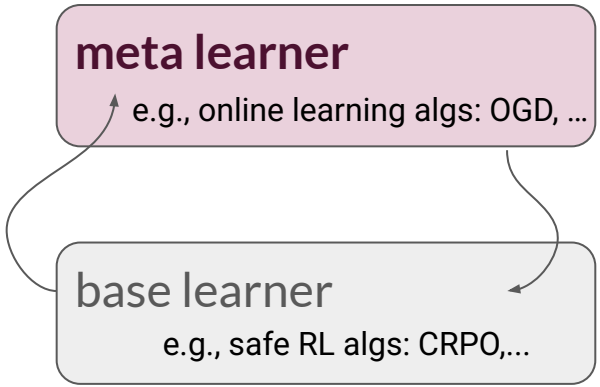


safety as performance
across complex, unknown,
extreme environments

closed world safety → **open world safety**

Case Study: Meta-Safe RL

Goal: to handle non-stationary environments and unknown scenarios by treating safety as a performance objective.



oversees/adapts base learner strategies

performs standard safe RL tasks

adaptive safety bound

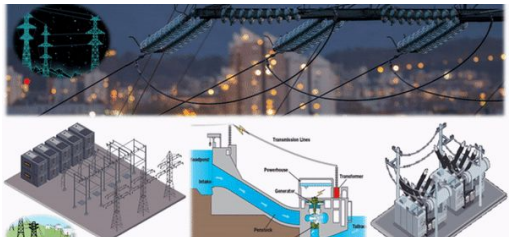
$$\mathcal{O}\left(\frac{\text{Env. variations}}{\sqrt{\#\text{ steps}}\#\text{ past CMDPs}}\right)$$

the more challenges (# CMDPs) encountered, the faster to be safe in future situations (reduced suboptimality gap)

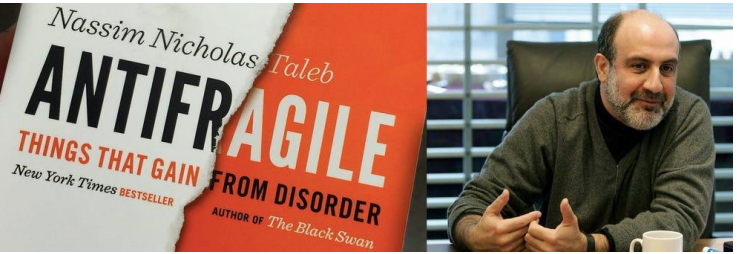
[Khattar et al., 2023]

♠**adaptive safety** mechanisms ♠**hierarchical structure** ♠first theoretical guarantee

critical load restoration to quickly restore the grid in blackouts [ul Abdeen et al., 2024], future applications in robotics, healthcare, autonomous driving, financial markets.

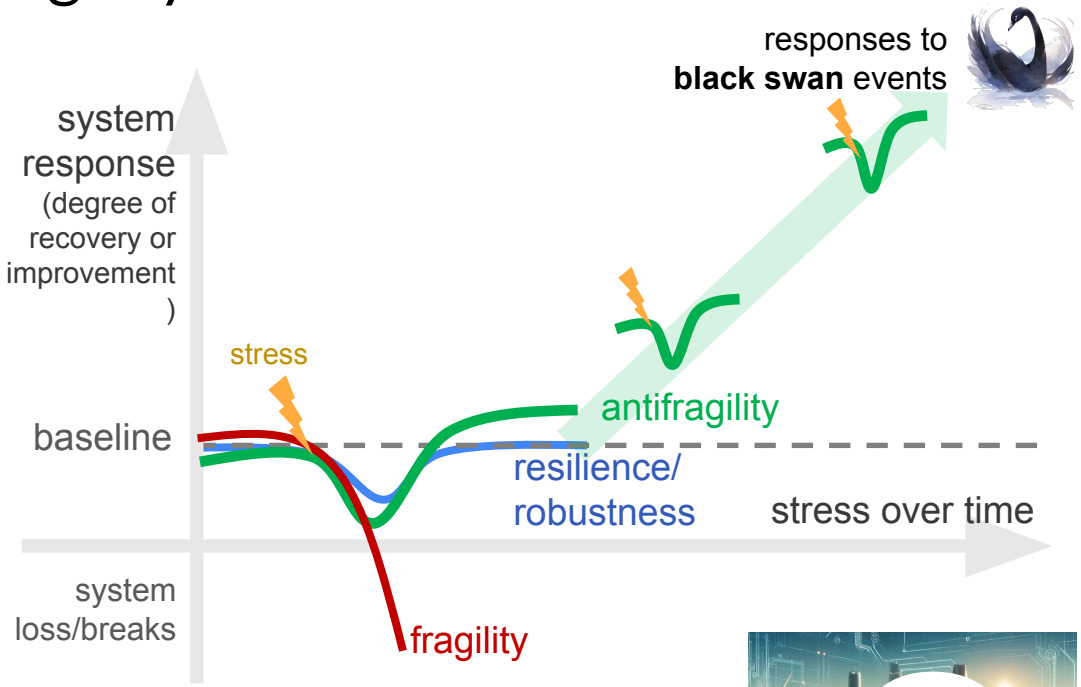


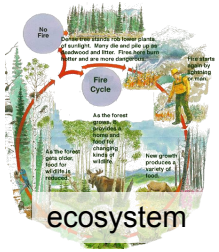
The Revelation of Antifragility



> some things **benefit from shocks**; they **thrive** and **grow** when exposed to **volatility, randomness, disorder,** and **stressors** and *love* adventure, risk, and uncertainty.

antifragility is *beyond resilience or robustness*. the resilient resists shocks and stays the same; the **antifragile gets better**.





immune
system

Towards Antifragile Power System

Preparing for Black Swans: The Antifragility Imperative for Machine Learning

Ming Jin*

*these research directions highlight the path towards creating **unusually safe** and **antifragile** decision-making systems capable of handling Black Swan events.*

nonstationary online learning

adapting to changing environments

improving with volatility

partially observable MDPs

capturing hidden factors and uncertainties

inferring and adapting to unobserved dynamics

safe exploration & control

learning under safety constraints

expanding boundaries safely and purposefully

continual/ lifelong learning

building on past knowledge, balancing stability and plasticity

adapting over time

meta-learning

learning to learn, adapting learning strategies

rapid adaptation

foundation models

on-the-fly adaptation via in-context learning

leveraging memory and retrieval for test-time adaptation

quality-diversity & multi-obj. learning

maintaining diverse solutions, handling conflicts

dynamic repertoire of strategies

adversarial ML

revealing vulnerabilities, ensuring robustness to extremes

strengthening through “vaccination”

The Future of Safe RL

Reinforcement Learning
enabling AI to learn optimal decision-making through interaction & feedback

safety assurance
guaranteeing reliable and responsible AI behavior in all scenarios.

Power system impact
deploying AI to solve complex challenges in power systems

call to action: envision, innovate, collaborate, and shape the future of safe, intelligent systems.

Questions? Contact Ming Jin (jinming@vt.edu)

Safe RL Seminar

Acknowledgements:

NSF · Commonwealth Cyber Initiative · C3.ai · 4-VA · U.S. DEPARTMENT OF ENERGY