Reinforcement learning for emergency voltage control and blackout prevention in electrical grids

Adrien Bolland
CMAP, École Polytechnique, 12 November 2026

Reinforcement learning for emergency voltage control

An electrical grid is a system that is actively controlled. After some events the controlled system may become unstable. We use reinforcement learning to avoid these situations.

This material results from a shared project with Louis Wehenkel and Paul Mangold.



Louis Wehenkel



Paul Mangold

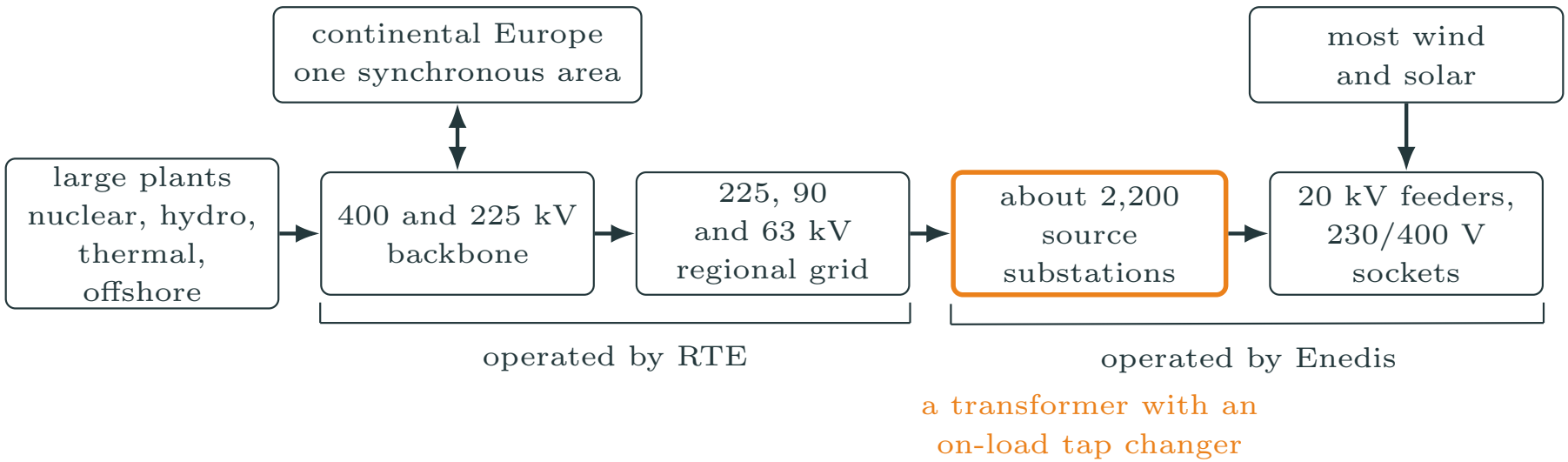
1. How electricity reaches a home
2. Controlling the on-load tap changer
3. Our problem statement
4. The DRL toolbox
5. Results
6. Open problems and take-away

Personal websites: https://people.montefiore.uliege.be/lwh/ and https://pmangold.fr/.

How electricity reaches a home

From the plant to the socket

The voltage is stepped down from the 400 kV backbone to the socket; power follows the impedances, not the contracts.



[1] CRE. Présentation des réseaux d'électricité, https://www.cre.fr/.

Market and operator responsibilities

Producers and retailers settle the schedule and RTE ensures the grid delivers it.

	months ahead	the day before	the same day	real time
markets	forward and bilateral contracts	day-ahead auction	intraday trading close to real time	imbalances settled
operator	outage planning, grid studies	reserves bought	N-1 studies, preventive actions	controllers, dispatchers

A contract does not guarantee delivery:

- The schedule is protected against the loss of any single element.
- The operator holds reserves and has access to control means.
- Some plants regulate their voltage based on a reference.

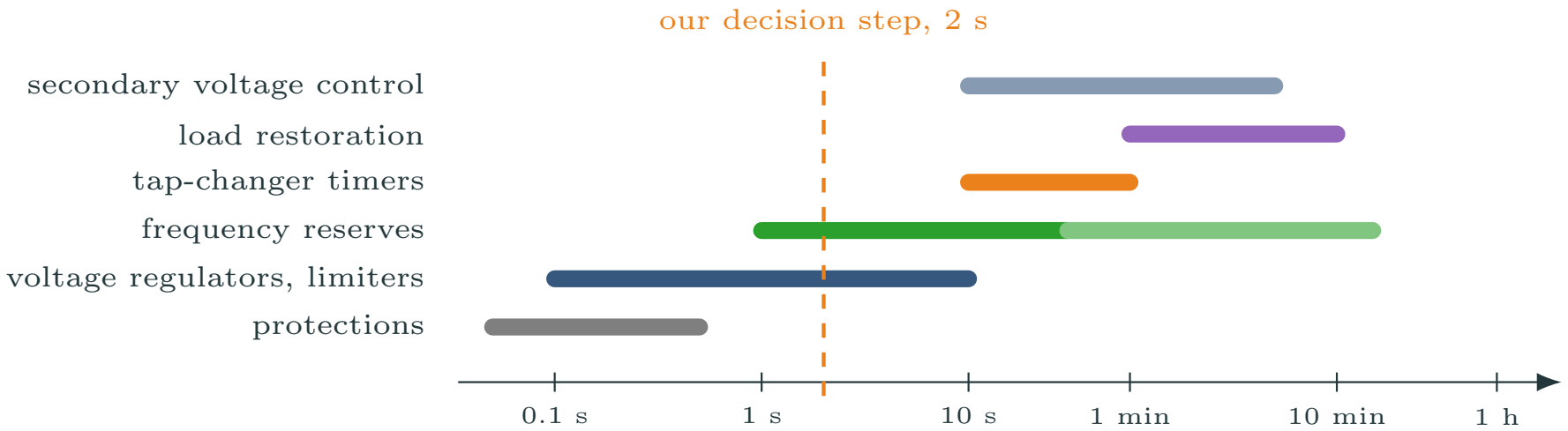
---

[1] European Commission, 1 Oct. 2025. [2] CRE. Délibération 2025-153. [3] RTE. Réserve rapide.
[4] Regulation (EU) 2016/631. [5] Regulation (EU) 2017/1485, Art. 19(3). [6] RTE. Safety report 2025, p. 39.

In real time the automatic controllers never stop

Frequency is global: all machines share the same frequency.

Voltage is local: reactive power must be produced at places where voltage drops.

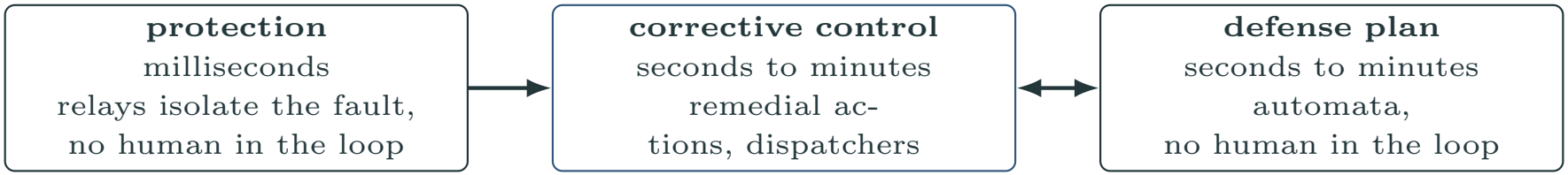


[1] Kundur, P. et al. Definition and classification of power system stability. IEEE TPWRS 2004.

[2] RTE. Safety report 2025, pp. 7, 39. [3] Fonteneau, R. Rapport de stage RTE, 2026, p. 14.

When something trips: protection, correction, defense

Three layers act after an event.



The defense plan automatically protects the system against phenomena too fast for any human intervention but does not resolve the problem:

- regions that lose synchronism are separated; load is shed on low frequency;
- regional automata shed load against voltage collapse;
- a national scheme blocks the tap changers, in service since the collapse of 12 January 1987; its thresholds were crossed four times in 2025.

This talk is about a defense automaton for voltage issues.

[1] RTE. Safety report 2025, pp. 30, 35, 36, 45, 46.

And yet the lights went out

Two grids went down in minutes, one by low voltage, one by high voltage.

Athens, 12 July 2004

7:08	a 300 MW unit near Athens is lost
11:30	Athens voltages sink toward 90%
12:01	the unit returns; the decline stops
12:12	lost again; the decline accelerates
12:25	100 MW shedding ordered, 80 shed
12:37	a unit in Central Greece trips
12:39	the south collapses, 4500 MW lost

Iberian Peninsula, 28 April 2025

10:30	voltages brush the 435 kV limit
12:03	oscillations, again at 12:19; damping them raises the voltage
12:32:00	voltages start to rise across Spain
12:33:18	more than 2.5 GW of generation trips and load changes in 78 s
12:33:19	synchronism lost; defenses too late
12:33:21	cut from France; dark at 12:33:30

ENTSO-E on Iberia:

"the key phenomenon in the incident was the non-effectiveness of voltage control"

---

[1] Vournas et al. Postmortem analysis and data validation in the wake of the 2004 Athens blackout. TPWRS 2006.

[2] ENTSO-E. Grid incident in Spain and Portugal on 28 April 2025, final report 2026, pp. 9–15, 22, 87.

The common mechanism: voltage instability

Voltage stability is the ability of a power system to keep acceptable voltages at every bus after a disturbance. It is lost when restoring the operating point demands more power than the grid can supply or absorb.

Undervoltage instability

After a disturbance the loads restore the power they drew before it. If the grid can no longer deliver that power, each restoration step lowers the voltage further, until it collapses.

Overvoltage instability

After a disturbance the loads restore the power they drew before it. If the grid can no longer run so lightly loaded, because the reactive power its lines produce exceeds what its generators can absorb, each restoration step raises the voltage further, until units trip on overvoltage.

---

[1] Van Cutsem, T. and Vournas, C. Voltage stability of electric power systems. Kluwer 1998, ch. 2.

[2] Vournas, C. and Van Cutsem, T. On overvoltage instability. PSCC 2026.

Why it is a new problem: the reactive balance flips sign

The same grid now visits two voltage regimes within a day.

- Loaded lines consume reactive power (the voltage drops); lightly loaded lines and cables produce it (the voltage increases).
- 95% of wind and PV sits in distribution, unobserved by RTE.
- It absorbs reactive power in proportion to its output, whatever the voltage; a shutdown (e.g., on negative market prices) removes that absorption at once.
- RTE installed 14 new shunt reactors (973 Mvar) in 2025 to absorb the excess reactive power: the answer today is equipment.

One controller must handle both directions. The tap changer is an installed lever that costs little energy and acts on the restoration in both voltage regimes.

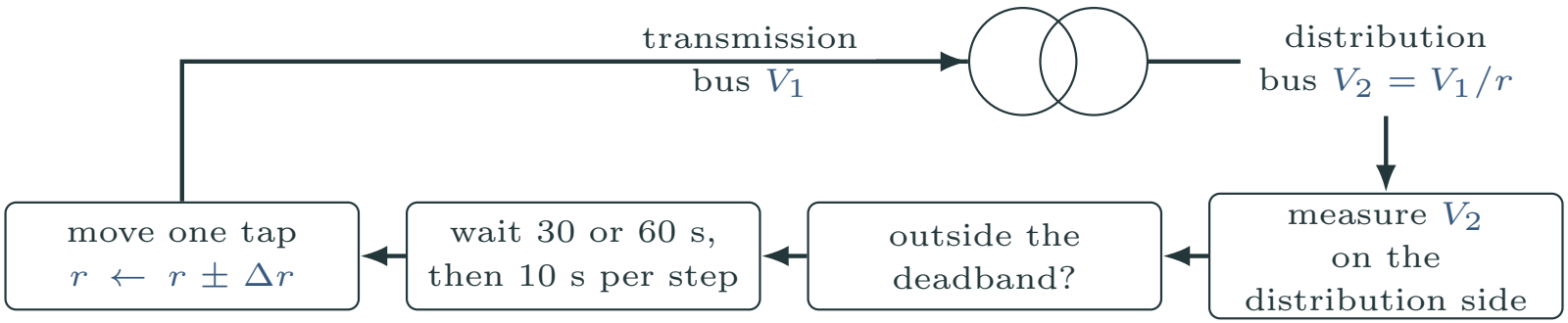
---

[1] Fonteneau, R. Rapport de stage RTE, 2026, pp. 8, 10–13. [2] RTE. Safety report 2025, pp. 36, 39–41.
[3] Vournas, C. and Van Cutsem, T. On overvoltage instability. PSCC 2026.

Controlling the on-load tap changer

What an on-load tap changer is and how it works

A transformer whose ratio moves under load. A local automaton drives it and sees only the distribution side.



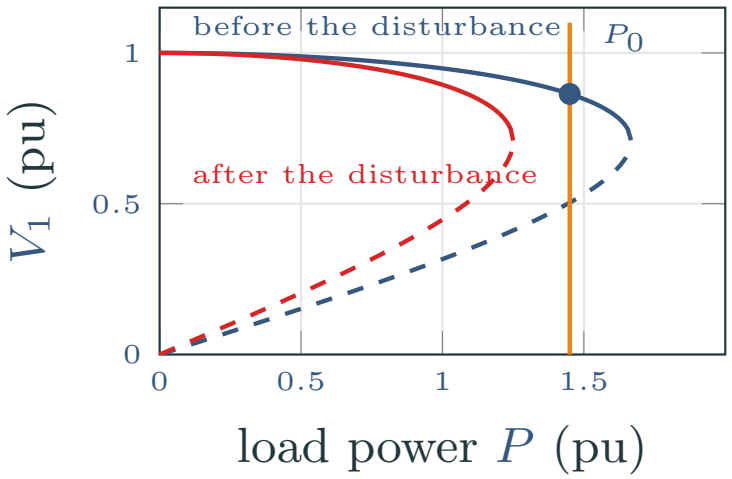
The automaton modifies the tap ratio r when the voltage is outside a deadband for longer than a delay. It keeps the distribution voltage in its band after any disturbance upstream, with no communication.

[1] Van Cutsem, T. and Vournas, C. Voltage stability of electric power systems. Kluwer 1998, ch. 4.

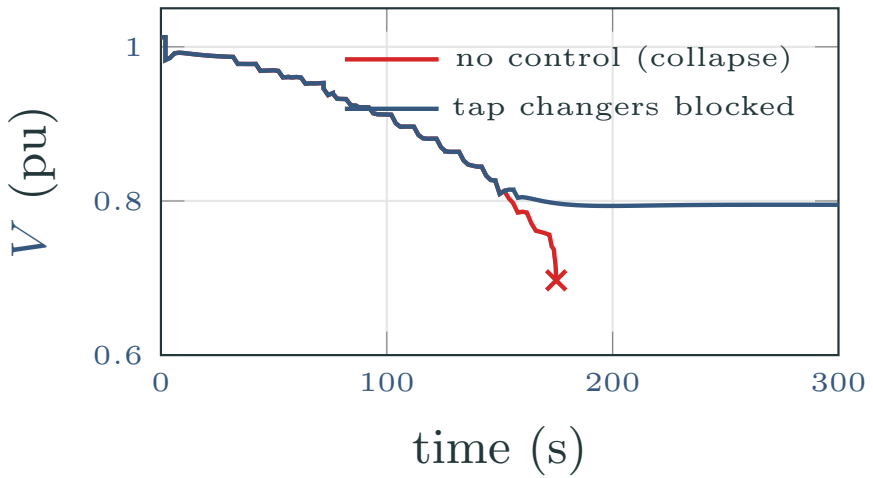
[2] Fonteneau, R. Rapport de stage RTE, 2026, p. 14.

The tap changer turns against the system

The tap changer decouples the distribution voltage from the transmission grid and makes the load consume constant power P_0.



Nordic system, voltage at bus 1041



Some disturbances may decrease the power the network can provide. If P_0 exceeds it, each tap step lowers V_1 further. Blocking holds r; the load becomes voltage-dependent again, and an equilibrium can exist again.

[1] Van Cutsem, T. and Vournas, C. Voltage stability of electric power systems. Kluwer 1998, ch. 4, 6.

[2] Traces: DynaWaltz use examples, Dynawo documentation, https://dynawo.github.io/.

Existing solutions

Existing methods fall into three families.

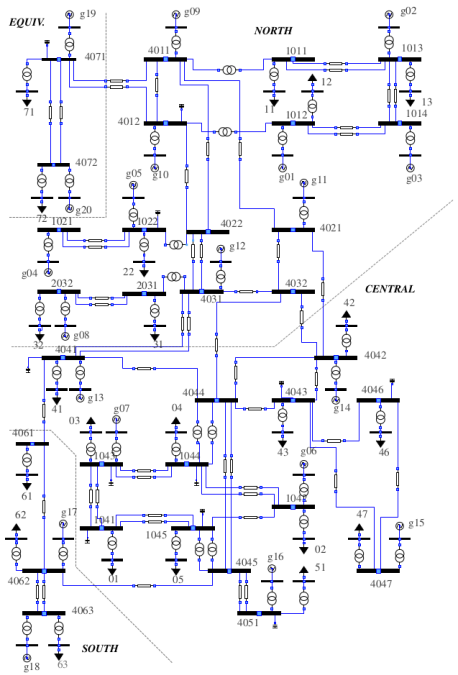
- Rule-based automata (1987–2009).
- Model predictive control (2002–2026).
- Reinforcement learning (2020–2025).

Rules: [1] Otomega 2003; [2] Vournas 2001; [3] Vournas 2008; [4] Capitanescu 2009. MPC: [5] Larsson 2002; [6] Wen 2004; [7] Beccuti 2010; [8] Glavic 2011; [9] Moradzadeh 2013; [10] Fonteneau 2026. Learning: [11] Huang 2020; [12] Hossain 2021; [13] Hagmar 2022; [14] Bi 2025; [15] Donon 2024.

Our problem statement

What we decide

Every 2 s, for each of 22 tap changers: idle, block or release.



The Nordic test system has 20 generators and 22 controllable tap changers.
Blocking freezes the ratio r; the distribution voltage then follows the transmission voltage.
Releasing hands r back to the automaton.

The two questions of the talk: which tap changers, and when.

---

[1] Van Cutsem, T. et al. Test systems for voltage stability studies. IEEE TPWRS 2020.

[2] Diagram: Dynawo documentation, https://dynawo.github.io/.

Step 1: the right dynamics

The slow power-system dynamics are hybrid differential-algebraic equations,

$$\dot{x} = f(x, y, z(k), a) \qquad 0 = g(x, y, z(k), a) \qquad z(k+1) = h(x, y, z(k), a)$$

where x collects the dynamic states (fluxes, controllers, limiters, restorative loads), y the network voltages and currents, z the tap positions, which move at sample times k, and a the commands.

The initial network is disturbed, and contingencies and load variations are drawn at random along the episode.

We use the DynaWaltz (Dynawo) simulator to generate trajectories.

---

[1] Guironnet, A. et al. Towards an open-source solution using Modelica (Dynawo). ISGT Europe 2018.

[2] Van Cutsem, T. and Vournas, C. Voltage stability of electric power systems. Kluwer 1998.

Step 2: the decision problem

A chance-constrained decentralized POMDP: n agents with observations o_i and commands $a_i \in \{\text{idle}, \text{block}, \text{release}\}$, contingencies and repairs in the transition kernel, a control cost $c(s, a)$. The problem is

$$\begin{aligned} \min_{\pi} \quad & \mathbb{E}^{\pi}\Big[\sum_{t=0}^{T} \gamma^t c(s_t, a_t)\Big] \\ \text{subject to} \quad & \mathbb{P}^{\pi}(\text{a voltage leaves the 10\% band}) \leq \alpha_e \\ & \mathbb{P}^{\pi}(\text{the system collapses}) \leq \alpha_c \end{aligned}$$

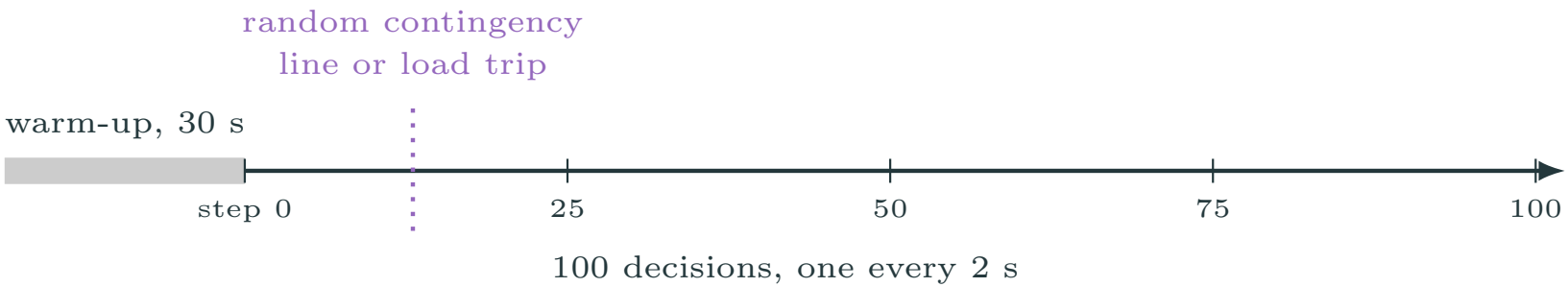
The constraints bound probabilities of first-hitting events, not cumulative costs.

---

[1] Bernstein et al. 2002. [2] Oliehoek and Amato 2016. [3] Altman 1999.

The instance we solve

Every agent sees the state, and the constraint violations are penalized with fixed weights. That instance is a multi-agent MDP (MMDP).



An episode is one disturbed network, 30 s of warm-up, then 100 decisions 2 s apart; a collapse ends it.

[1] Boutilier, C. Planning, learning and coordination in multiagent decision processes. TARK 1996.

What the formulation buys, and why learning it is hard

What the formulation brings:

- the timing and the subset of tap changers are learned;
- uncertainty is considered in the transition kernel;
- safety against under- and overvoltages is considered;
- execution needs no online optimization.

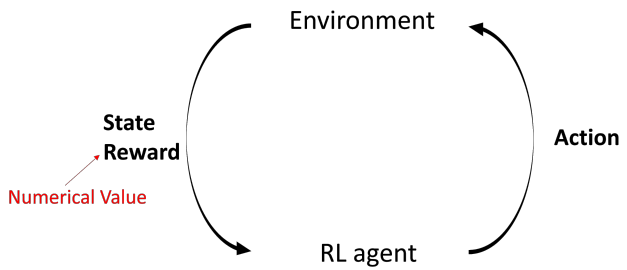
The cost of learning:

- each episode takes several seconds of simulation;
- optimal control under rare events;
- 3^{22} joint commands, with consequences a minute or two later.

The DRL toolbox

From formulation to learning

The simulator is the environment: we learn in simulation.



With learning we take the time offline, over many simulated episodes, to learn the policy properly; the policy is then fast at execution.

Soft actor-critic

Off-policy maximum-entropy soft actor-critic (SAC), in its discrete form:

$$v(s) = \sum_a \pi(a \mid s)\big[Q(s,a) - \beta \log \pi(a \mid s)\big]$$

where v is the soft value and β a temperature tuned to a target entropy.

- Sample efficiency: every transition goes to a replay buffer and is reused in many updates, whichever policy generated it.
- Rare events: the policy is a categorical distribution over the three commands, and the critic evaluates all of them at every state, so a rare but decisive command (block) is learned from the few times it was tried.
- The entropy target keeps probability on the rare commands during training.

[1] Haarnoja et al. Soft actor-critic. ICML 2018. [2] Christodoulou 2019. [3] Zhou et al. TMLR 2024.

Two policy classes

One command for all, or one per device.

Central

One categorical policy sends the same command to every tap changer. It blocks or releases the fleet at once.

Decentralized

One command per tap changer, conditioned on the state:

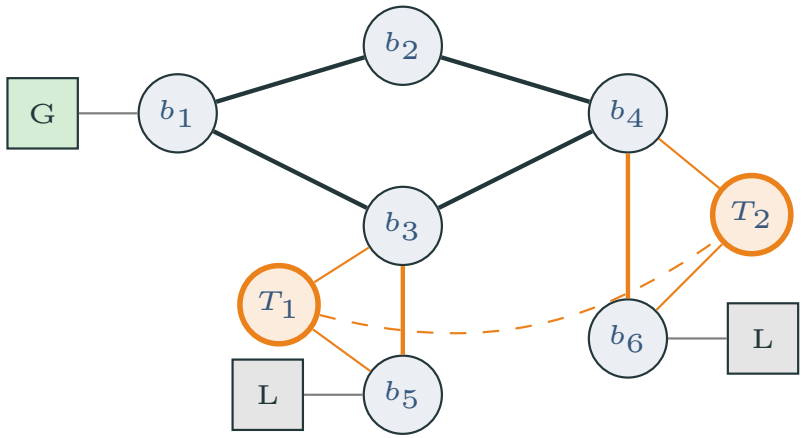
$$\pi(a \mid s; \theta) = \prod_i \pi_i(a_i \mid s; \theta)$$

For the MMDP, there exists such an optimal policy.

State representation

States are represented by typed graphs.

- buses: voltage and angle;
- generators: powers, speed, field voltage, regulator and limiter states;
- loads: active and reactive power;
- lines and transformers: resistance, reactance, flows, loadings;
- tap changers: position and lock flag.



orange: tap changers;
dashed: same zone, messages exchanged

The bus graph carries buses, generators, loads and branches, each with its own features; the tap-changer graph links each tap changer to its transformer buses and to the tap changers of its zone.

[1] Fey, M. and Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. 2019.

[2] Guironnet, A. et al. ISGT Europe 2018.

Graph neural networks

A graph neural network is an equivariant function: with node features $\mathbf{X}$ and adjacency $\mathbf{A}$, $F(\Pi\mathbf{X}, \Pi\mathbf{A}\Pi^\top) = \Pi F(\mathbf{X}, \mathbf{A})$ for any relabeling Π of the nodes.

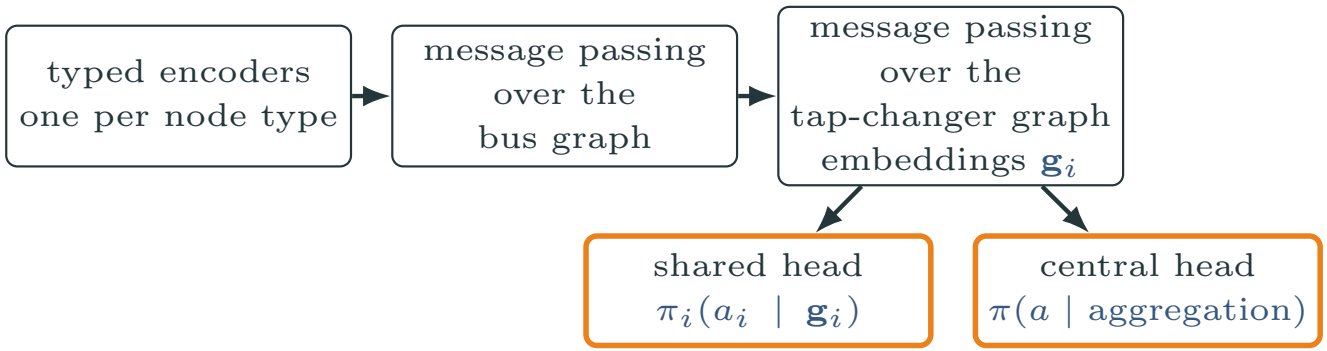
It is built from layers of message passing. Each node provides its node feature computed from its own feature and the sum of its neighbors',

$$\mathbf{h}_i \leftarrow \phi\Big(\mathbf{h}_i, \sum_{j \in \mathcal{N}(i)} \mathbf{h}_j\Big)$$

with the same function ϕ at every node. The sum does not depend on the order of the neighbors, so relabeling the nodes only relabels the output. After L layers a node sees its L-hop neighborhood.

[1] Bronstein et al. Geometric deep learning. arXiv 2021. [2] Corso et al. Nature Rev. Methods Primers 2024.

Our architecture

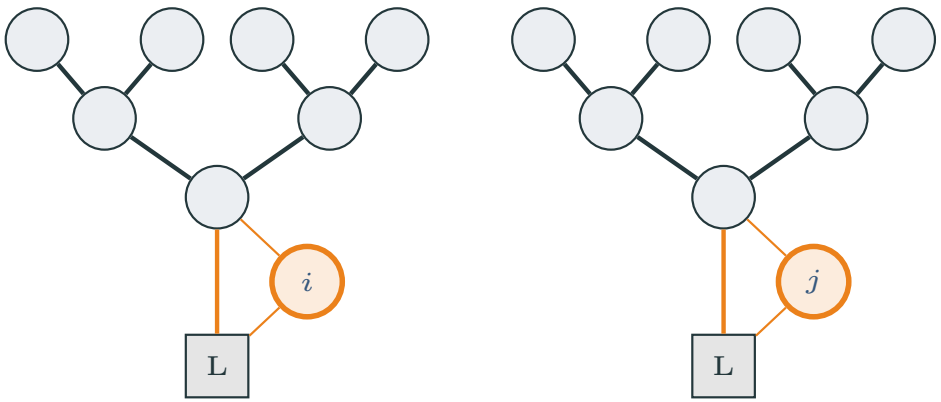


- One small network per node type encodes the raw features of the graph;
- Message passing over the bus graph gives one embedding per bus;
- Message passing over the tap-changer graph gives the embeddings $\mathbf{g}_i$;
- The policies use these embeddings as state representation.

[1] Donon et al. EPSR 2024. [2] Hossain et al. IEEE TPWRS 2021.

Positional encoding

Message passing cannot tell apart two nodes whose neighborhoods look alike.



tap changer i tap changer j

same neighborhood, so $\mathbf{g}_i \approx \mathbf{g}_j$

Each bus receives a signature,

$$\mathbf{p}_i = \left[(\mathbf{P})_{ii}, (\mathbf{P}^2)_{ii}, \ldots, (\mathbf{P}^{16})_{ii}\right]$$

where $\mathbf{P}$ is the transition matrix of a random walk that moves from bus to bus along the branches, and $(\mathbf{P}^k)_{ii}$ is the probability of being back at bus i after k steps.

[1] Dwivedi et al. Graph neural networks with learnable structural and positional representations. ICLR 2022.

Critic representation

The critic scores the joint command of all devices; it is only used during training.

- Value decomposition: a critic that is a sum of per-device terms, $Q_{\text{tot}}(s, a) = \sum_i Q_i(s, a_i)$, replaces a function of 3^{22} joint commands.
- Asymmetric methods: the critic may see more than the policy.

---

[1] Oliehoek and Amato 2016. [2] Sunehag et al. 2018. [3] Rashid et al., QMIX, 2018. [4] Zhang et al., FOP, 2021.

[5] Baisero and Amato 2022. [6] Lyu et al. 2023. [7] Ebi et al. 2025.

Safety and the reward

Safe reinforcement learning addresses constraints with Lagrangian multipliers, barrier functions, safety critics that estimate hitting probabilities, and risk measures such as the conditional value-at-risk.

We penalize the constraint violations with fixed weights:

$$R_t = -\kappa \cdot \text{effort} - \lambda \cdot \text{security violation penalties}$$

The effort depends on blocked tap changers; penalties are given for voltage excursions and collapse. A collapse ends the episode.

The weights κ and λ trade blocking against voltage quality.

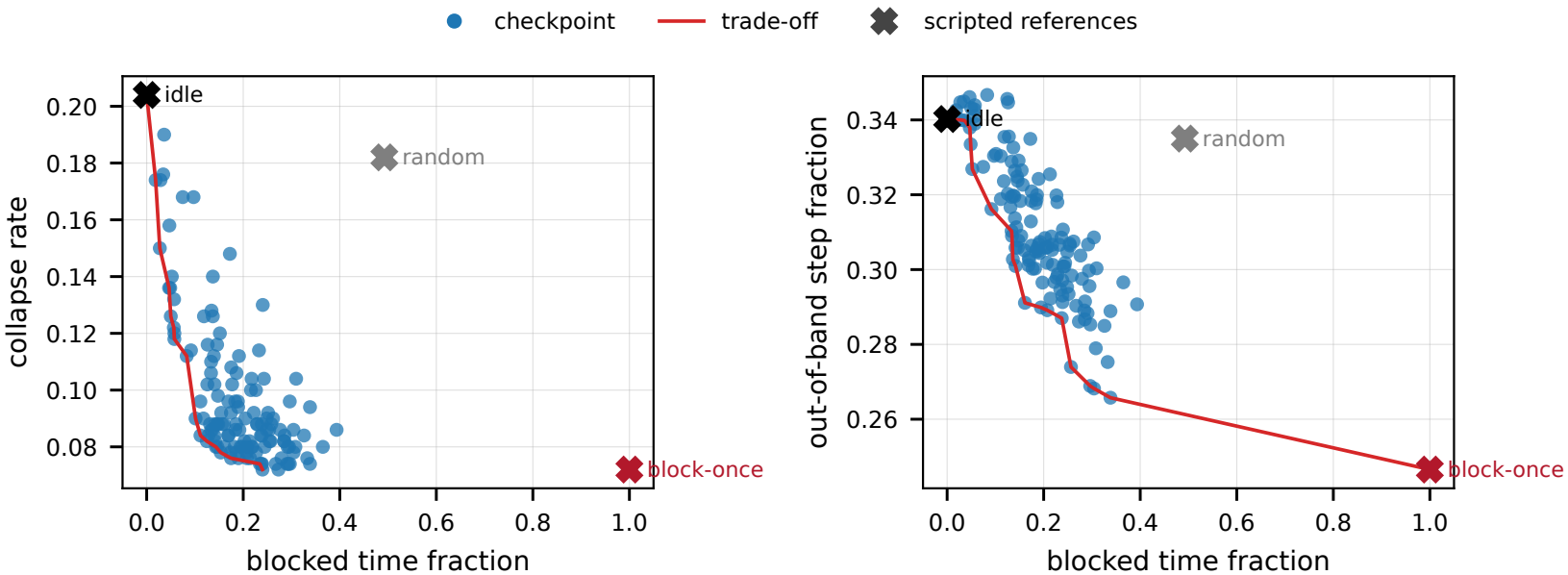
---

[1] Altman 1999. [2] Tessler et al. 2019. [3] Stooke et al. 2020. [4] Rockafellar and Uryasev 2000. [5] Tamar et al. 2015. [6] Chow et al. 2017. [7] Yang et al. 2021. [8] Vu et al. 2021. [9] Thananjeyan et al. 2021.

Results

Physical trade-offs

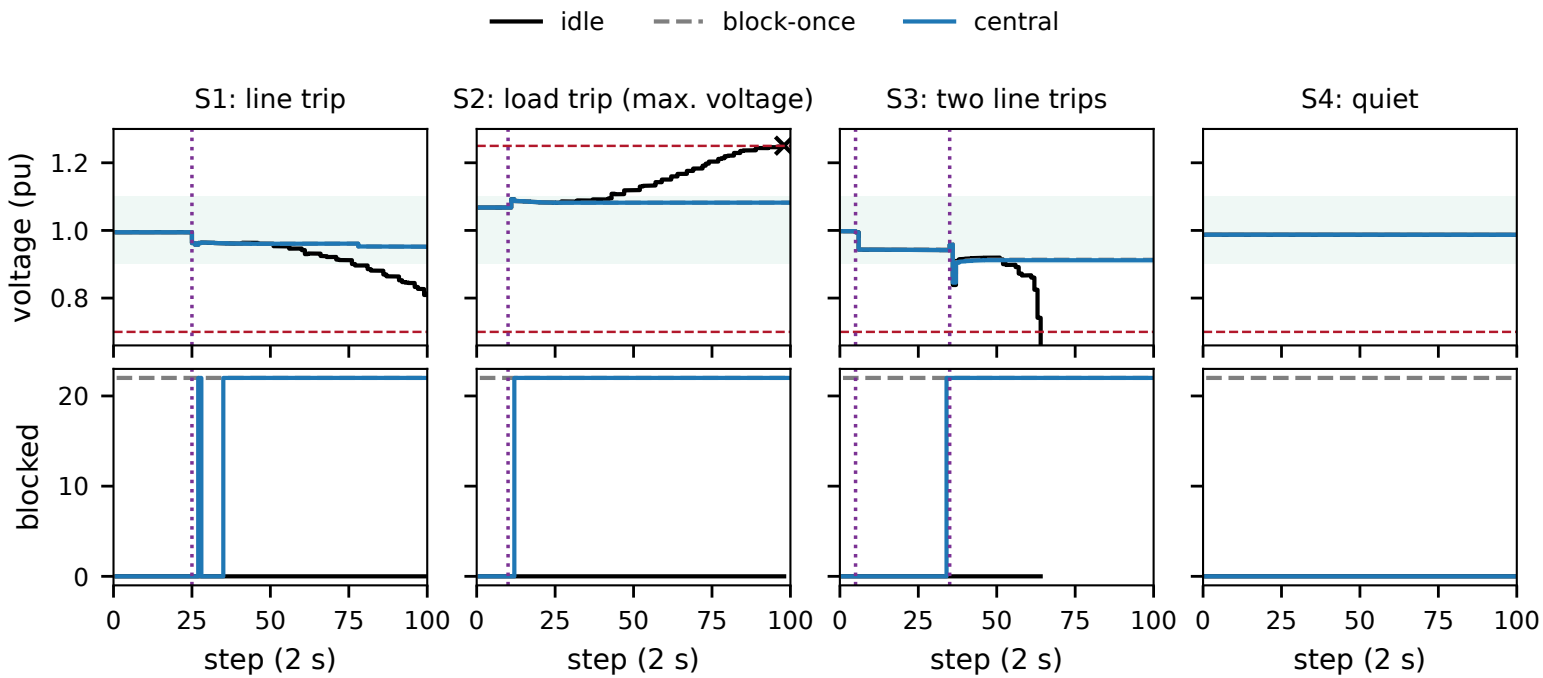
The weights move the controller along a security-versus-cost trade-off.



Each point is a checkpoint of a reward study with other weights, evaluated on 500 episodes: collapse rate and time out of the band against the fraction of time its tap changers are blocked.

The central controller in time

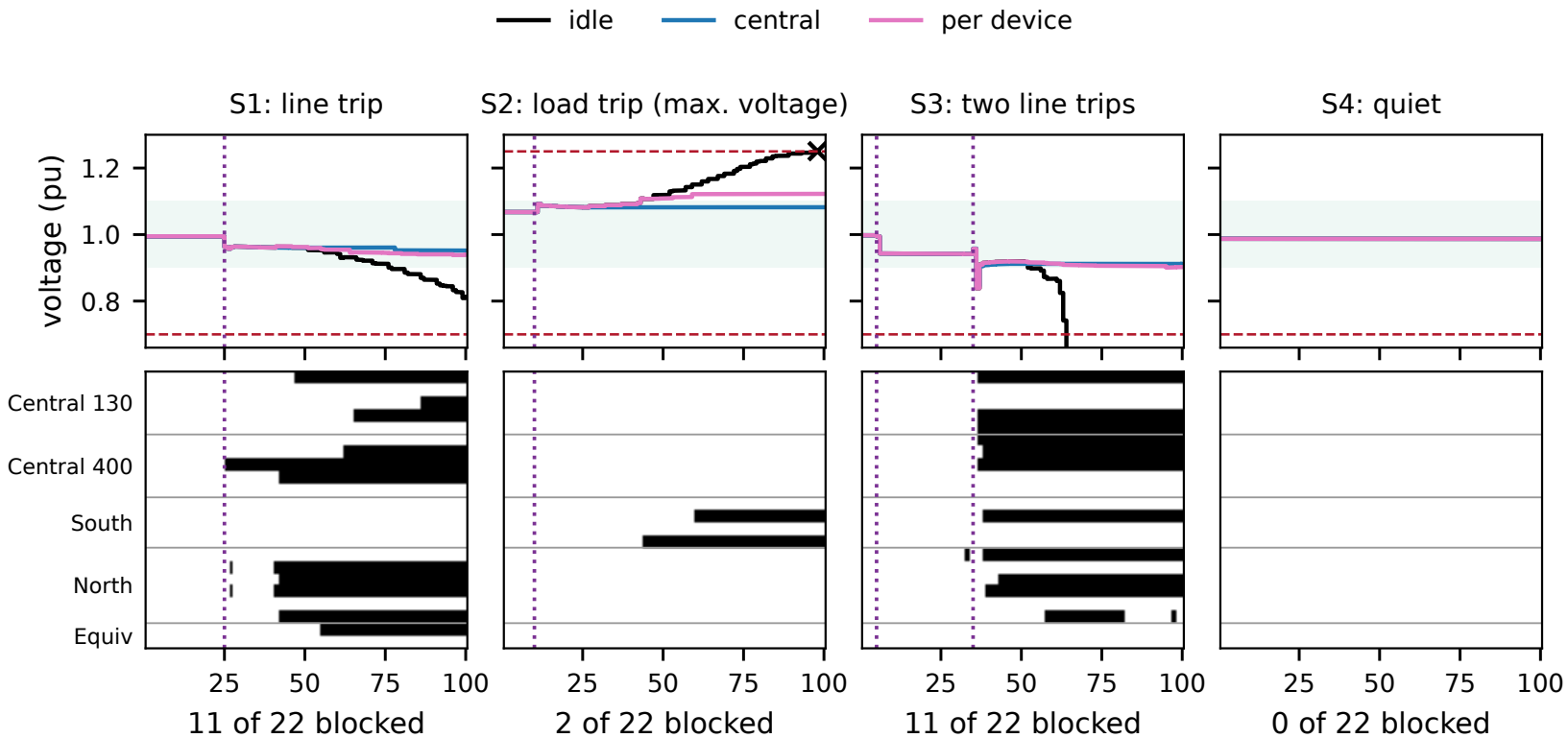
One command for all: the controller waits for the event, then blocks the whole fleet of tap changers.



On four held-out cases it blocks two steps after the trip of S1, releases, and holds from step 35; two steps after the trip of S2; one step before the second trip of S3; never on S4. Over 500 episodes and four seeds: collapse rate 5.6%, a block on 22% of the steps.

One command per tap changer

The decentralized class blocks fewer tap changers for an individualized control.



Over the same 500 episodes: collapse rate 5.3% against 5.6%. It holds 14% of the fleet blocked on average against 22%, but some tap changer is blocked on 52% of the steps.

Open problems and take-away

Open problems and take-away

Open problems

- Policy classes for combinatorial action spaces and the cost of expressivity.
- Chance constraints with primal-dual methods.
- Partial observability and communication between agents.
- Rare events evaluation with few seeds.
- Simplified dynamics and action models.
- Guarantees for the deployed policies.

Take-away

Voltage instability is back, and now in both directions. A learned controller can cut the collapse rate while accounting for uncertainty, deciding per tap changer, and being deployed at high frequency.