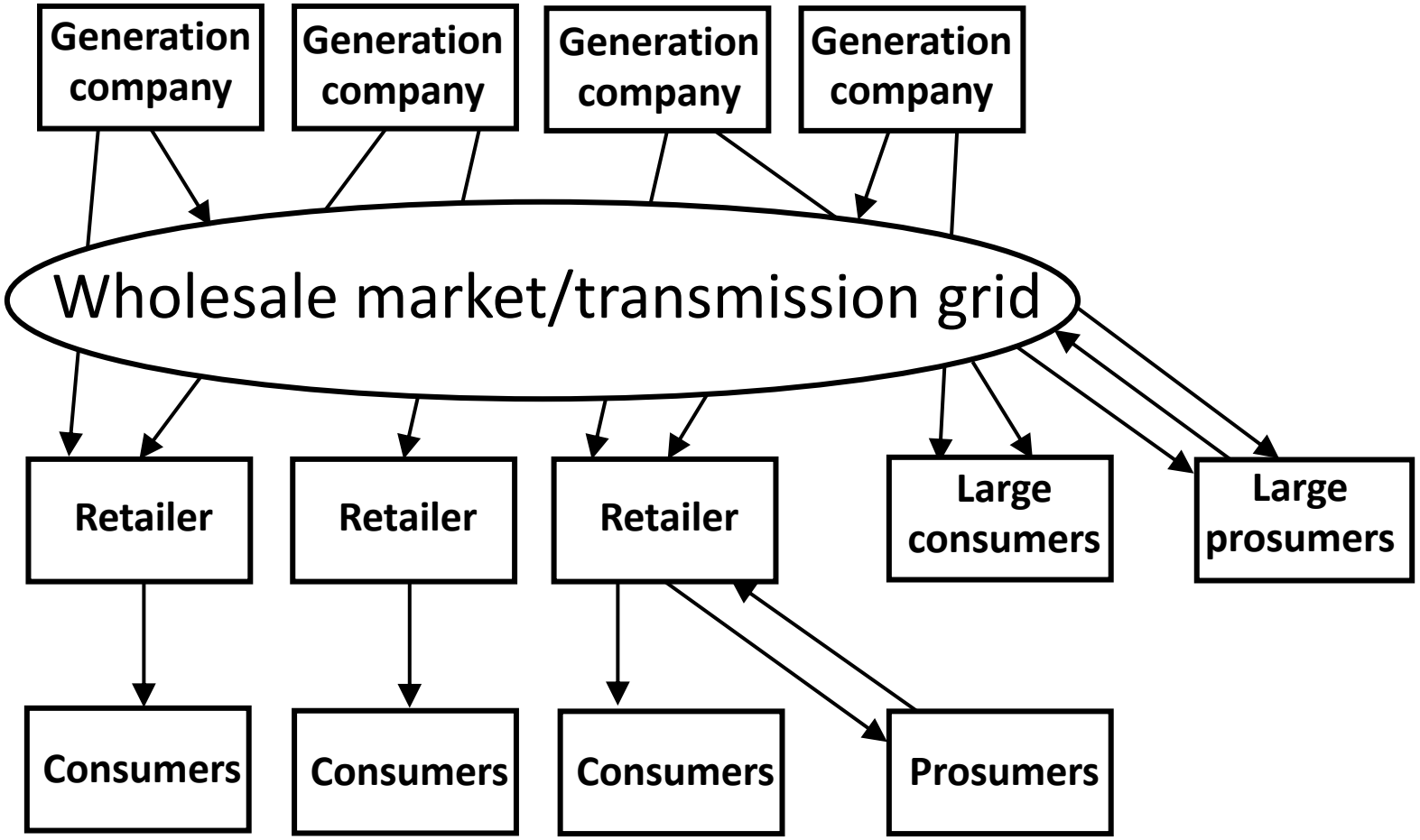

Reinforcement learning for
electrical markets and the energy transition
Prof. Damien ERNST - CSO Haulogy
LIÈGE université
haulogy
The Energy IT Company

The retail model for electricity markets

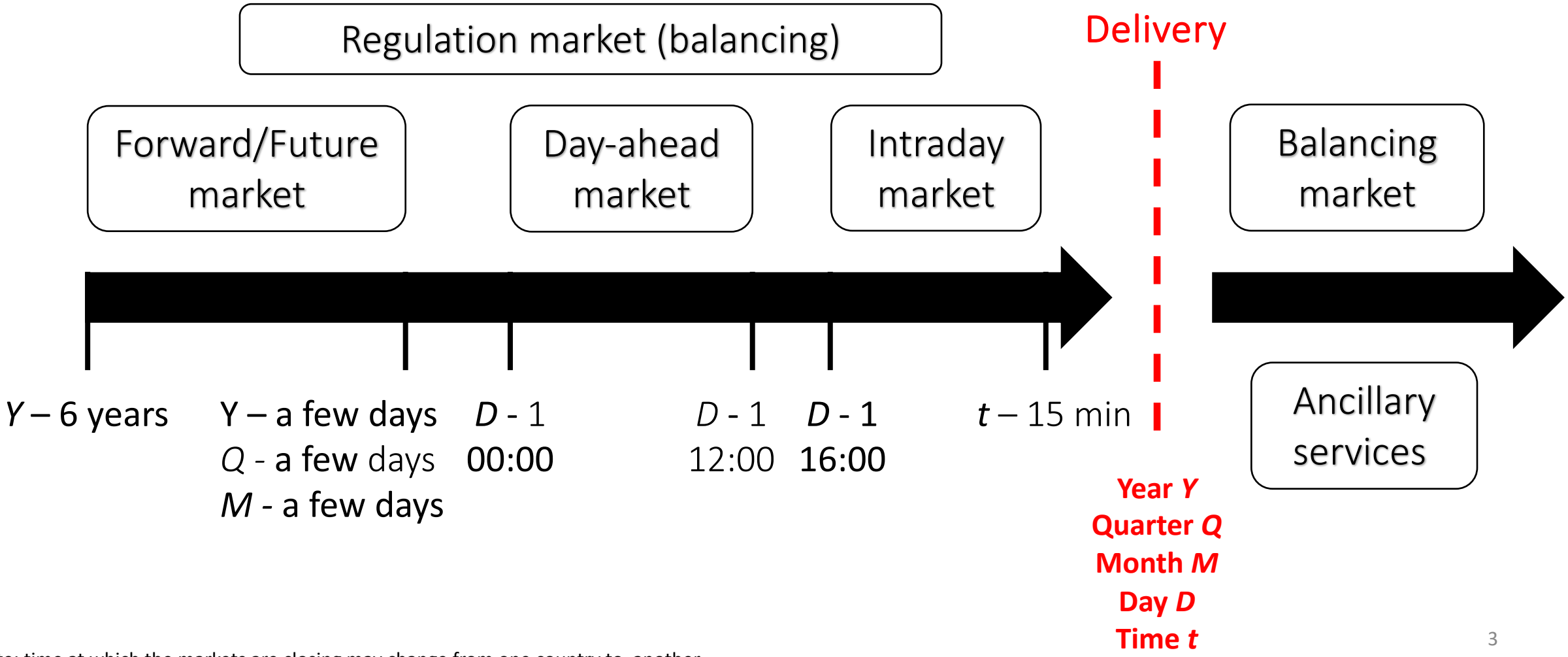


Dominant model in Europe for more than ten years.

When you buy electricity, you buy a quantity of electrical energy that will be delivered over a specific time period. In the EU, this is typically 15 or 30 minutes.

The current market structure is evolving with new players appearing (renewable energy communities, etc.) and consumer-centric models becoming popular.

Electricity markets – Overview



Note: time at which the markets are closing may change from one country to another.

Prices for monthly, quarterly and yearly products on the forward markets: an example

Actual prices (Last quotation: 2/05/2023, end of day)

Month	Today	Yesterday	% D/D	5 days high	5 days low	Year high	Year low
Jun	98,78 €	100,60 €	-1,81%	104,39 €	98,78 €	141,44 €	98,78 €
Jul	101,98 €	101,96 €	+0,02%	108,88 €	101,96 €	134,03 €	101,96 €
Aug	101,73 €	102,23 €	-0,49%	102,23 €	101,73 €	102,23 €	101,73 €

Quarter	Today	Yesterday	% D/D	5 days high	5 days low	Year high	Year low
Q3-23	106,58 €	106,91 €	-0,31%	112,46 €	106,58 €	195,92 €	106,58 €
Q4-23	143,84 €	144,24 €	-0,27%	145,25 €	142,10 €	216,94 €	135,18 €
Q1-24	161,32 €	160,83 €	+0,31%	161,32 €	154,13 €	203,54 €	140,96 €
Q2-24	121,76 €	120,68 €	+0,89%	121,76 €	117,69 €	129,35 €	117,69 €

Year	Today	Yesterday	% D/D	5 days high	5 days low	Year high	Year low	Historic high	Historic low
2024	138,20 €	138,15 €	+0,03%	138,20 €	132,53 €	176,71 €	114,72 €	396,50 €	46,64 €
2025	131,59 €	131,05 €	+0,41%	131,59 €	127,59 €	150,72 €	115,08 €	267,50 €	71,23 €
2026	114,75 €	113,75 €	+0,88%	114,75 €	112,12 €	128,21 €	101,91 €	128,21 €	101,91 €

Why deploy RL for interacting with electricity markets?

Reinforcement learning (RL) refers to a set of techniques that enable one to solve a complex ***stochastic sequential optimal decision-making problem*** solely from the information generated through interaction with the environment defining the dynamics of the problem and its reward function.

Reason #1: RL techniques target the maximization of a numerical signal (monetisation of the RL agent behaviour) which is convenient when you want to maximize profits.

Reason #2: Retailers, generators and consumers interacting with electricity markets are interacting with a sequence of markets. In these markets, they must buy/sell the exact amount of power they consume/produce for every market period or be exposed to the balancing market. The many loads/generation devices/batteries whose flexibility can be exploited to maximize gains also lead to a time-coupling of the decisions. Hence the problem is naturally sequential.

***Reason #3*:** Problems are highly stochastic and it is often very difficult to model the environment in an appropriate way.

Ambitious research objective: *To design an RL agent able to simultaneously interact near-optimally with all these markets, whatever the player (producer, retailer, prosumer, etc.).*

RL for interacting with the continuous intraday market

The continuous intraday market (CID) allows market participants to **adjust their positions** through **bilateral contracts**. It is an opportunity for producers and consumers to make **last-minute adjustments** and balance their positions **closer to real time**. This electricity market authorises **continuous trading**, meaning that a trade is executed as soon as two orders match.

We focus on the problem of an RL agent that controls a battery while interacting at the same time with the intraday market to maximize its profits. The agent has a constraint to be balanced for every market period. That leads to a decoupling with the balancing market.

Based on the following paper:

A Deep Reinforcement Learning Framework for Continuous Intraday Market Bidding
BOUKAS, IOANNIS; ERNST, DAMIEN; THÉATE, THIBAUT; BOLLAND, ADRIEN et al. In *Machine Learning* (2021)

CID Market Design

Table 1. Order Book for Q_1 and time slot 00:00-00:15

i	**Side**	v **[MW]**	p **[€/MWh]**	
4	"Sell"	6.25	36.3	
2	"Sell"	2.35	34.5	⟵ ask
1	"Buy"	3.15	33.8	⟵ bid
3	"Buy"	1.125	29.3	
5	"Buy"	2.5	15.9	

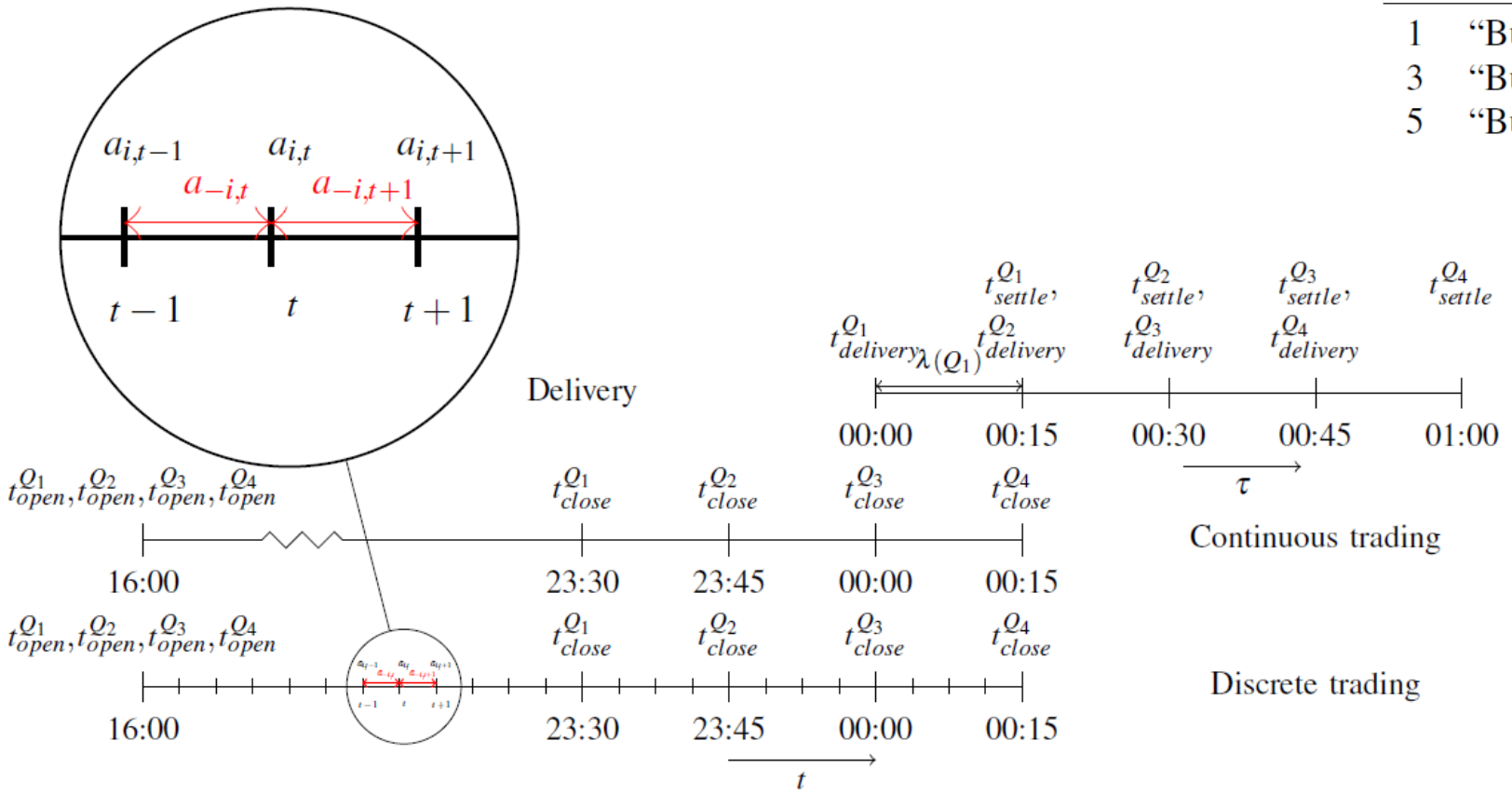
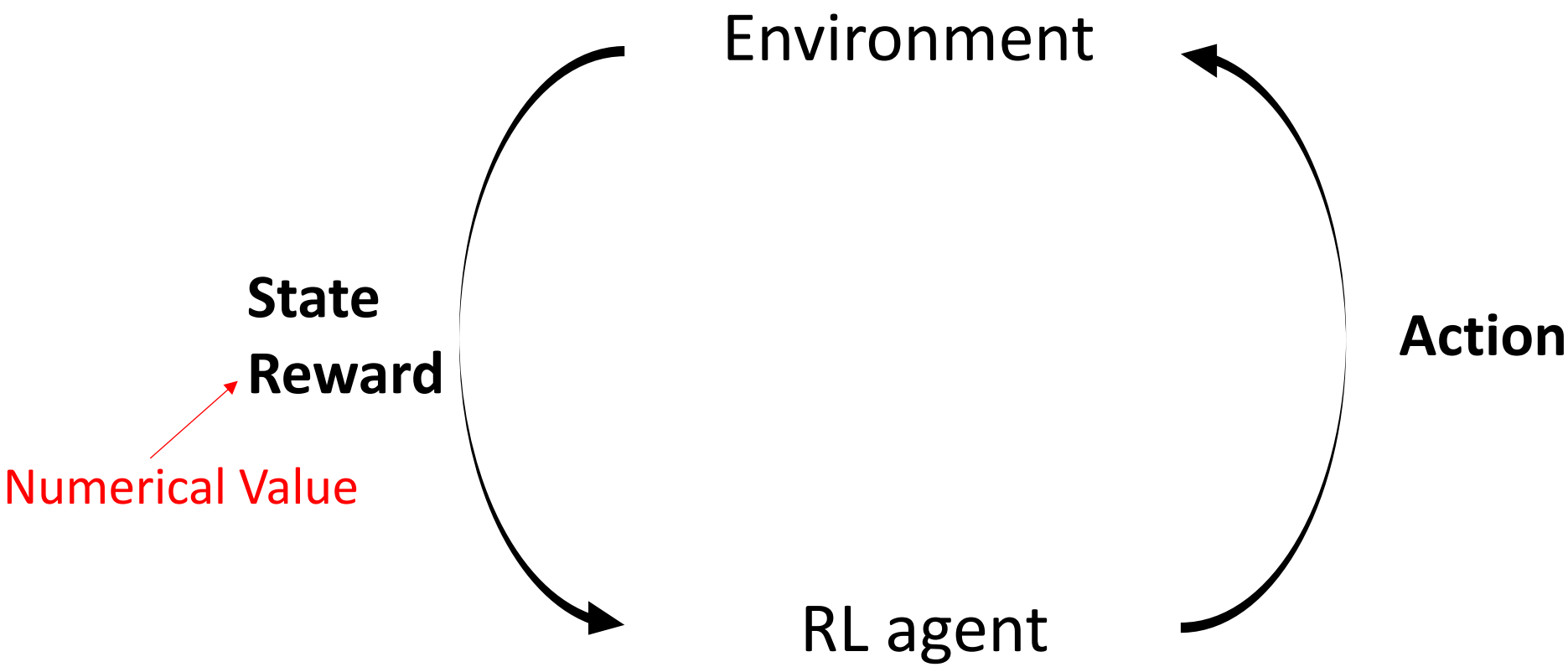


Fig. 1: Trading (continuous and discrete) and delivery timelines for products Q_1 to Q_4

We assume that the agent can only grab (partially or totally) orders in the market or change the setting of the battery at the beginning of every market period.

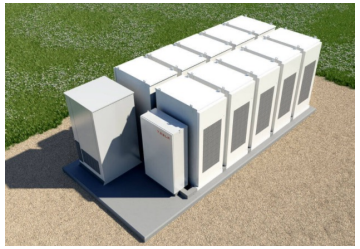
The Generic RL Scheme



RL scheme applied to CID trading with a battery

State: (i) the battery state of charge (ii) everything you know about the market – knowledge assumed to be limited to the contents of the order book.

Reward: the money made during the market period.



+ the energy market

The learning agent that controls the battery and determines the interaction with the CID market.

Action: the power generated by the battery for the next market period and the market orders that have been grabbed.

Reducing the complexity of the problem

State-space: comprised of what exists between the space of all possible order books and the set of possible states for the battery. *Solution*: reduction of the order book by representing it through simple features.

Action space: Number of actions at one instant equal to the number of elements in the order book plus one, and the power setting for the battery. *Solution*: defining two high-level actions namely, Idle or Trade. The Trade action grabs orders and computes all future settings of the battery to maximize profits under the assumption that no new orders will be posted in the order book and that the system should be balanced all the time.

Adversarial setting: When you play with markets, you are essentially in an adversarial setting, but we have used an RL algorithm designed for a non-adversarial setting for training our agent. This is a reasonable solution if you are not a too big a player in the market.

The way the learning has been done

Trajectories have been generated by an RL agent interacting with the ‘training part’ of the order book and the battery. An Epsilon-greedy policy is used.

The **Fitted Q Iteration** (FQI) algorithm has been used for extracting high-performing policies from the trajectories. Due to the finite time setting of the problem, one Q-function has been learned for each time step *t.*

Approximation architecture made of one layer of LSTM neurons and five fully connected feedd-forward layers.

The quality of the policy is assessed on the ‘testing parts’ of the order book. A comparison is done with the so-called rolling intrinsic (RI) policy that applies the Trade action at every instant *t.*

	FQI policy		RI policy
	V (€)	r (%)	V (€)
mean	667.9	3.8	645.2
min	153.7	-26.7	181.6
25%	490.9	-0.7	476.5
50%	649.9	4.0	620.8
75%	814.1	9.9	753.3
max	1661	40.9	1398
sum	102937	-	98071

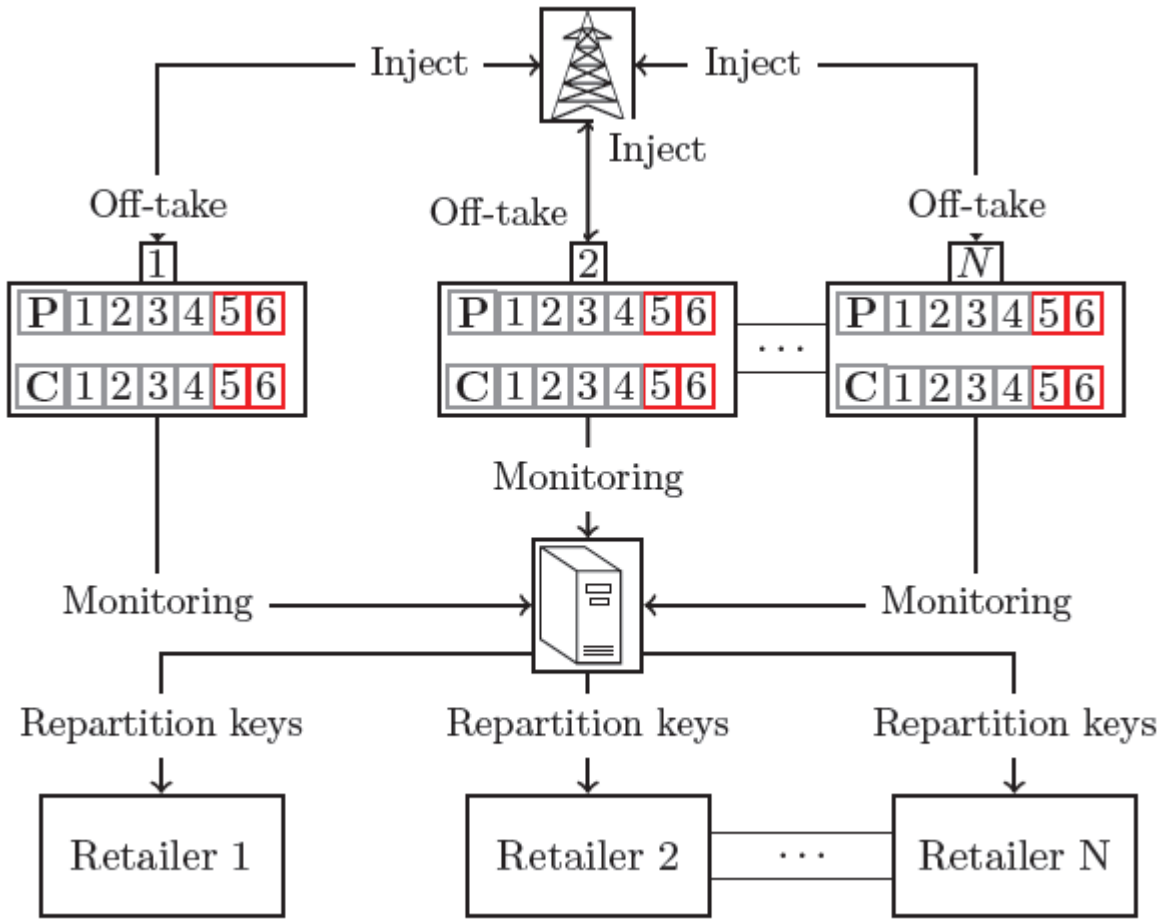
V is the profit per testing day and *r* is the probitability ratio with respect to the RI policy.
Testing would reflect the exact real-life performances of the FQI policy if its behavior would indeed not influence the trading behavior of other agents.

Renewable Energy Communities

The new Renewable Energy Directive 2018/2001 introduces the concept of Renewable Energy Communities (REC)

An REC is a goup of members connected to the main utility grid (often the distribution network) who consume electricity and/or produce renewable electricity. Batteries and/or flexible loads may be present in the REC.

Periodically, surplus electricity production is reallocated by the REC manager to the consumers through a system of **repartition keys**. These determine the fraction of the surplus of electricity production to be reallocated to each member. They define the way electricity is shared among the different members.



Each member of the REC is associated with a consumption meter (C) and a production meter (P). At each discrete time step *t*, these meters are incremented by the consumption and the production of each member during the time interval (t; t + 1], respectively. These meters are monitored at each end of a metering period by an **Energy Management System (EMS)** that controls the flexibility devices. The EMS computes optimal repartition keys with the values read from all the meters of the REC members and sends them to their respective retailers. Retailers periodically (e.g., monthly) compute the electricity bills for each REC member based on the repartition keys and sends them to their respective customers.

Controlling the REC

The problem of controlling the REC to maximize gains can be formalized as an optimal-sequential decision-making problem. More information about this rather difficult formalization:

Optimal Control of Renewable Energy Communities with Controllable Assets
AITTAHAR, SAMY; MANUEL DE VILLENA MILLAN, MIGUEL; CASTRONOVO, MICHAËL; BOUKAS, IOANNIS et al.

It is very important to jointly optimize the repartition keys and the controllable assets.

Reinforcement learning techniques combined with load/generation prediction techniques can be used.

Sizing complex energy systems

Sizing an energy system refers to the process of computation of the investments that will maximize benefits when the system is operated in a (near-)optimal way.

It has become more and more difficult to size modern energy systems such as a Renewable Energy Community because of the complexity of the environment they are associated with (different types of markets, dynamic grid fees, new devices such as e.g. (autonomous) electrical vehicles, power-to-X devices appearing, etc.).

The solution we propose: the use of newly developed **reinforcement learning algorithms for jointly optimizing the environment and the control policies**.

Jointly designing and controlling a system

There are two paradigms for solving such problems:

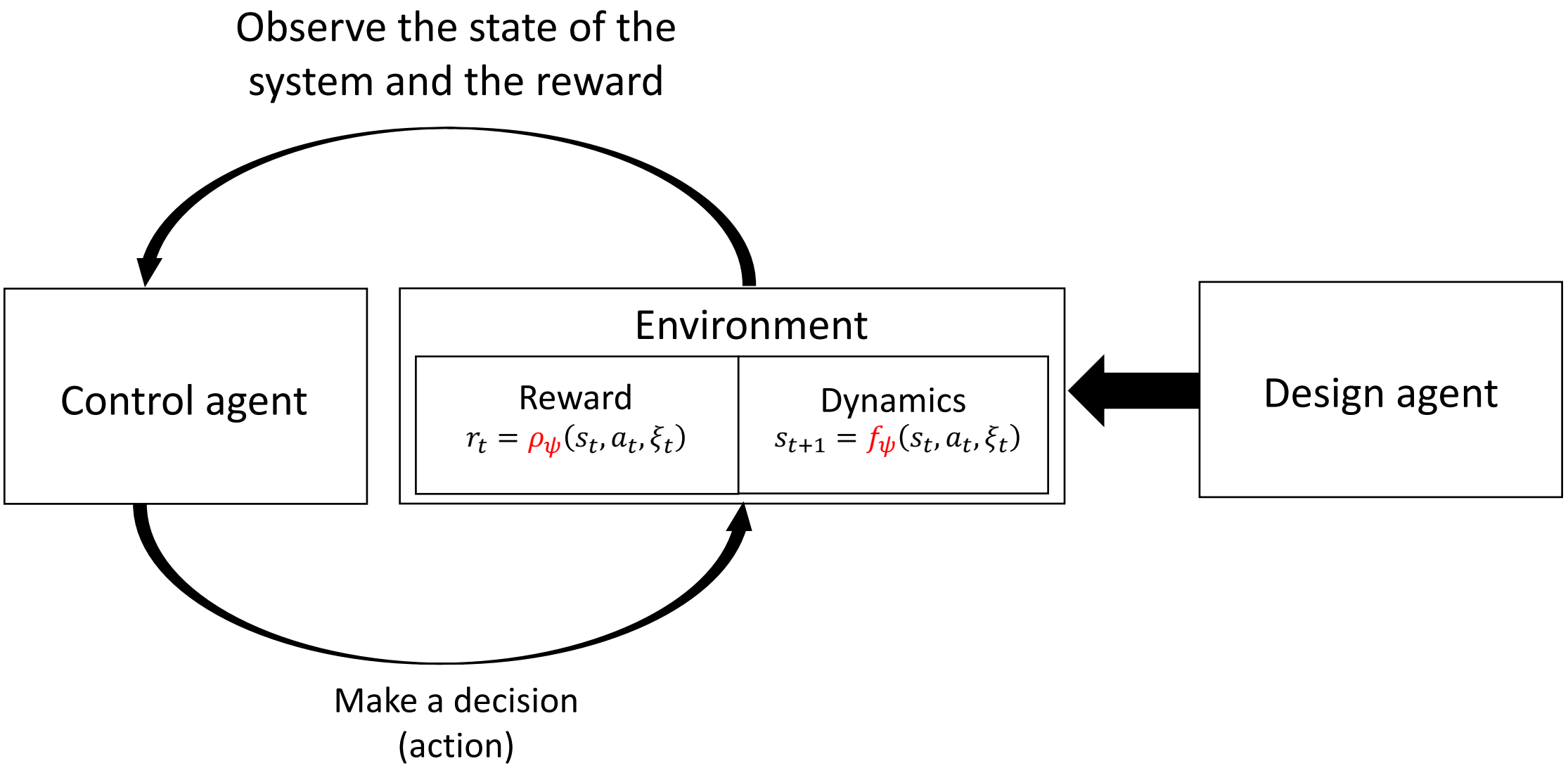
(i) Mathematical programming → Does not account properly for stochasticity, non-linearities, etc.

(ii) Sample-based (RL) techniques → Do not scale well to complex problems. Example the Joint Optimization of Design and Control (JODC) algorithm.

We propose a hybrid algorithm called **Direct Environment and Policy Search (DEPS)**. The algorithm lies in between reinforcement learning and model-based optimization. It combines sample-based techniques with the computation of derivatives of the dynamics and the reward function for efficient solution search. More information:

Jointly Learning Environments and Control Policies with Projected Stochastic Gradient Ascent BOLLAND, ADRIEN; BOUKAS, IOANNIS; BERGER, MATHIAS; ERNST, DAMIEN.

Direct Environment and Policy Search (DEPS)



1. Start with random environment parameters ψ .
2. Define a control agent whose policy is modelled using a Deep Neural Network with θ as parameters.
3. Initialise the parameters θ at random
4. Repeat iteratively
    1. Simulate trajectories in the system with the control agent
    2. From the trajectories, the equations of the system dynamics and of the reward function compute the gradients $\mathrm{V}(\psi, \theta) = E\left\{\sum r_t\right\}$ with respect to θ and ψ.
    3. Update the parameters:

$$\theta \leftarrow \theta + \alpha \nabla_{\theta} \mathrm{V}(\psi, \theta)$$
$$\psi \leftarrow \psi + \alpha \nabla_{\psi} \mathrm{V}(\psi, \theta)$$

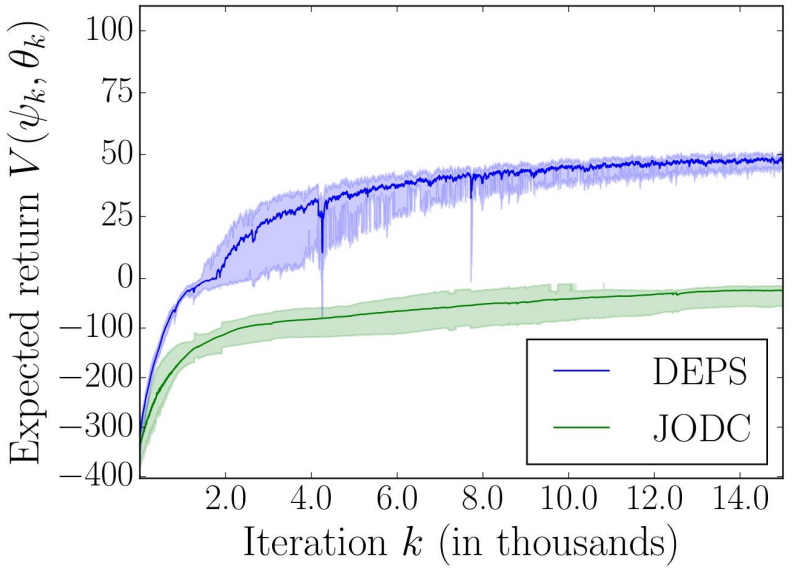
Joint design and control of a microgrid

Design choices: relates to the installed capacity of PV panels, the battery capacity and the maximal output of the diesel generator.

Control policy: sets the power output of the battery and the generator.

Objective: minimizing the sum of the investment and operational costs.

Results: DEPS outperforms the baseline JODC. It gives better overall performance and is a more stable algorithm.

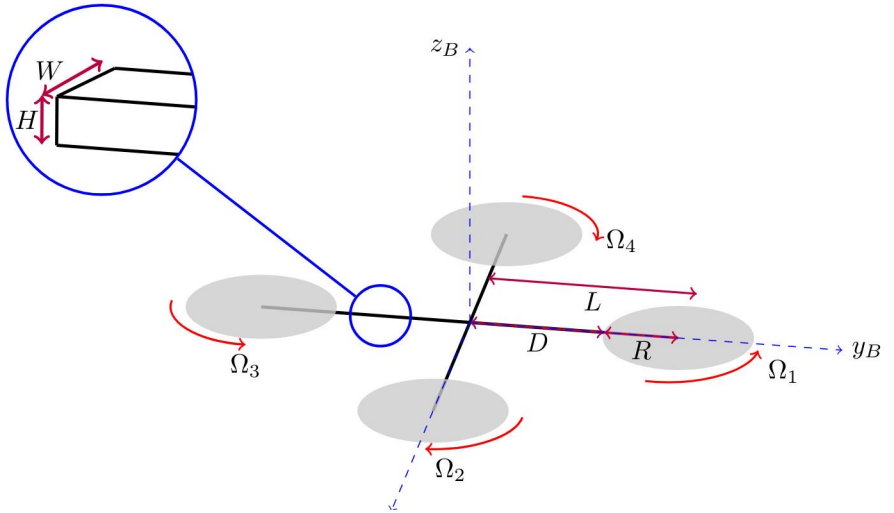


Joint design and control of a drone

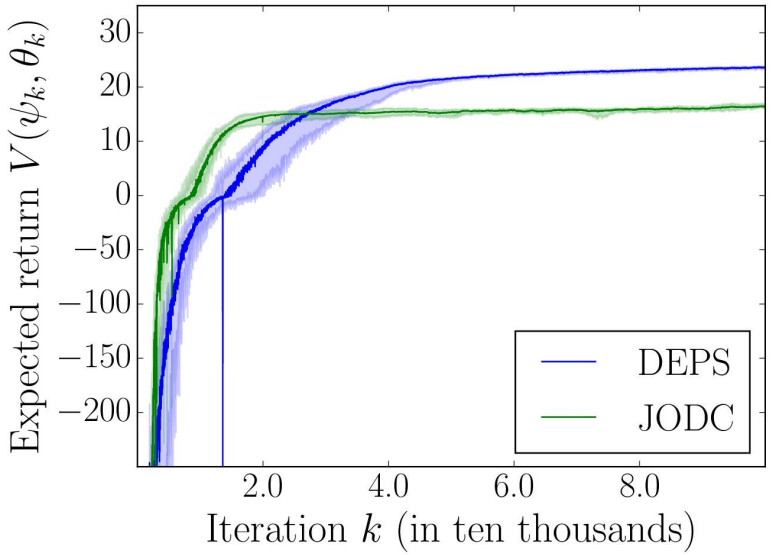
Design choices: the morphology of the drone defined by the parameters *H*, *W*, *D* and *R*.

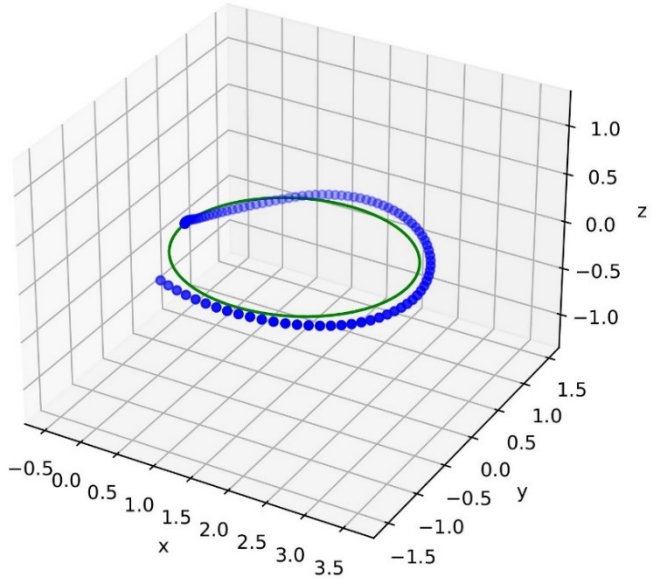
Control policy: sets the speed of the four propellers.

Objective: flying as accurately as possible on an ellipse.



Results:





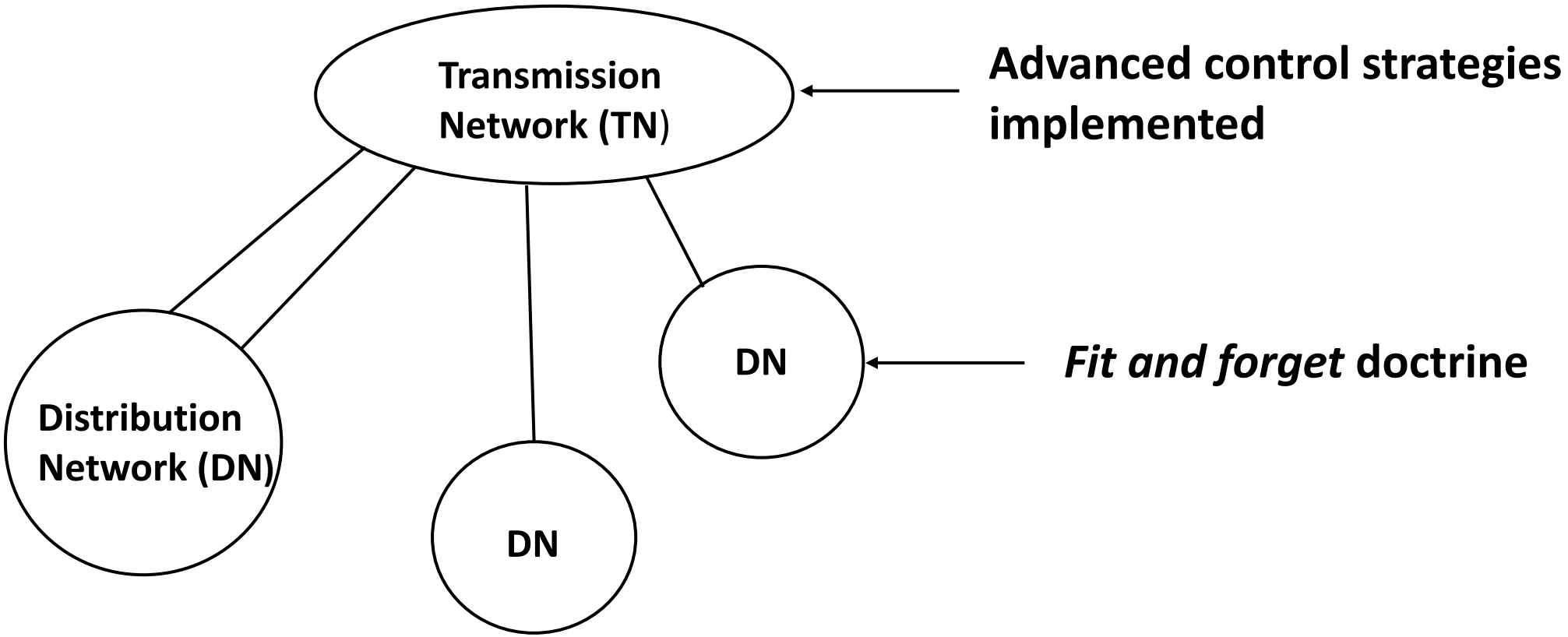
Let us not forget the electrical grid!

Up to now we have focused on control/design problems relevant for agents interacting with the electrical grid and energy markets.

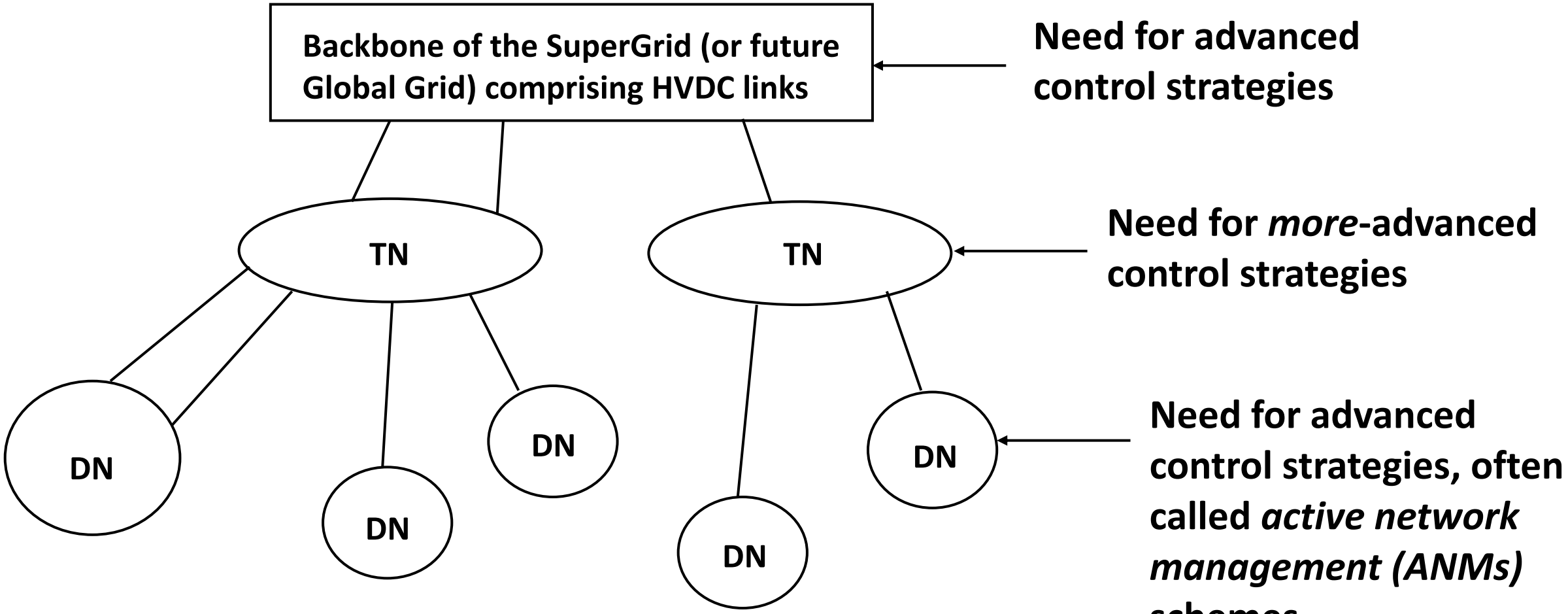
However, grid operators are themselves facing complex control problems related to energy transition and modifications of the control structure of the electrical grid!

Reinforcement learning could help grid operators to design efficient control strategies.

The grid before the quest for renewable energy



The grid as it is becoming



Gym-ANM

Observation: Decision-making problems related to the electrical grid have drawn less attention from the RL community than other fields over the last decade (e.g., video games, robotics, autonomous driving). We believe this is because:

1. RL research relies heavily on the availability of software-based simulators to model the real world during training.
2. Power systems are hard to simulate without extensive knowledge in power system engineering.
3. This makes it difficult for RL researchers to start working on power system decision-making problems.

Our solution: Provide an open-source software library (**Gym-ANM**) with which:

1. Power system experts can design new environments (*tasks*) that model ***Active Network Management (ANM)*** problems.
2. RL researchers can train and evaluate the performance of their agents on existing tasks (created in **1.**).

More information: **GYM-ANM: Reinforcement learning environments for active network management tasks in electricity distribution systems** HENRY, ROBIN; ERNST, DAMIEN in *Energy and AI* (2021), 5

Key Features of Gym-ANM

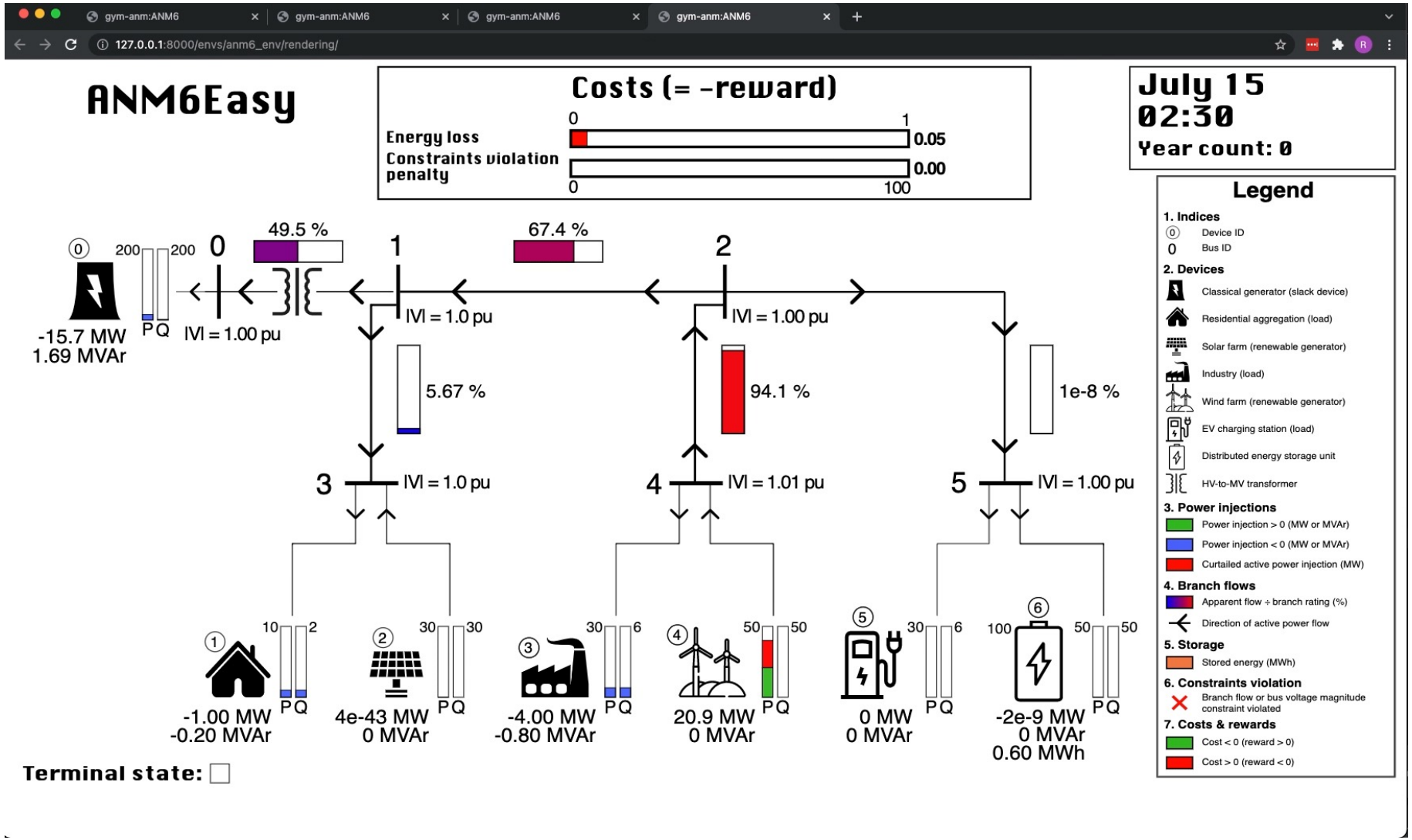
The environments (*tasks*) constructed with Gym-ANM follow the OpenAI Gym framework, which a major portion of the RL community is already using.

The different customisable components of Gym-ANM makes it suitable for modelling a wide range of ANM tasks:

- Example: simple ones for educational purposes.
- Example: complex ones for advanced research.

Once designed, each environment can be used as a "blackbox" → RL researchers need not be concerned with its implementation details.

Video example



Video also available on Youtube: https://youtu.be/D8kGH94kavY

References

A Deep Reinforcement Learning Framework for Continuous Intraday Market Bidding BOUKAS, IOANNIS; ERNST, DAMIEN; THÉATE, THIBAUT; BOLLAND, ADRIEN et al.

Optimal Control of Renewable Energy Communities with Controllable Assets
AITTAHAR, SAMY; MANUEL DE VILLENA MILLAN, MIGUEL; CASTRONOVO, MICHAËL; BOUKAS, IOANNIS et al.

Jointly Learning Environments and Control Policies with Projected Stochastic Gradient Ascent
BOLLAND, ADRIEN; BOUKAS, IOANNIS; BERGER, MATHIAS; ERNST, DAMIEN. GYM-ANM: Reinforcement learning environments for active network management tasks in electricity distribution systems HENRY, ROBIN; ERNST, DAMIEN.

GitHub repository (3K+ downloads after 6 months of release) for Gym-ANM: https://github.com/robinhenry/gym-anm and the online documentation: https://gym-anm.readthedocs.io/en/latest/index.html


Thank you for coming!
My twitter account:
@DamienERNST1