Efficient Management of Day-Ahead Energy Markets via Multi-Agent Reinforcement Learning - a Hybrid Model Case Study

Anonymous authors
Paper under double-blind review

Abstract

This study examines the optimization of day-ahead hybrid electricity markets. The shift from centralized systems to public-private models introduces many challenges, including the introduction of independent market players and *renewable energy sources* (RESs). A formal model of market participants' behavior is developed, and a *multi-agent reinforcement learning* (MARL) framework is proposed to optimize system operator strategies, incorporating dynamic pricing and dispatch scheduling to reduce operational costs, ensure stability, and align market incentives. A new and adaptable simulation environment, compatible with state-of-the-art methods, is presented. Evaluations in increasingly complex settings demonstrate the efficacy of our framework in managing the complexities of modern electricity markets.

1 Introduction

This work addresses the day-ahead optimization of an electricity market[1] undergoing significant structural transformation. Historically centralized and government-controlled, the increasing integration of *renewable energy sources* (RESs) and the advancements in data collection technologies are transitioning the market into a complex public-private hybrid model. This presents substantial challenges and the need to deal with a highly uncertain operational and regulatory environment Zhu et al. (2023).

To demonstrate some of the challenges involved in managing current energy systems, consider a day-ahead market in which the ***independent system operator*** **(ISO)** aims to optimize electricity generation based on forecasted demand, generation costs, and grid constraints. The resulting decisions, made 24 hours in advance, specify the amount of electricity to be produced, the prices, and the allocation of reserve capacity, i.e., the ability to generate additional power at short notice, often at high environmental costs, in the event of generation failures or unexpected demand surges.

Adapting the day-ahead market to today's energy systems requires accounting for the variability and limited controllability of increasingly heterogeneous ***grid-edge agents***, denoted hereon as **GEAgents**, particularly those with local generation and storage capabilities. For example, a household with a photovoltaic (PV) unit and a battery can autonomously optimize its energy storage policy, learning when to store energy, when to consume it, and when to trade with the grid to maximize economic benefits. While such behavior may improve individual utility, it introduces significant uncertainty into aggregate demand forecasts and can destabilize the system, especially under sudden shifts in consumption or generation patterns. At the same time, these distributed resources can enhance efficiency and resilience by shaving peaks, supplying energy, and reducing the amount of centrally dispatched generation required.

[1] For anonymity reasons, the specific market under consideration is not disclosed.

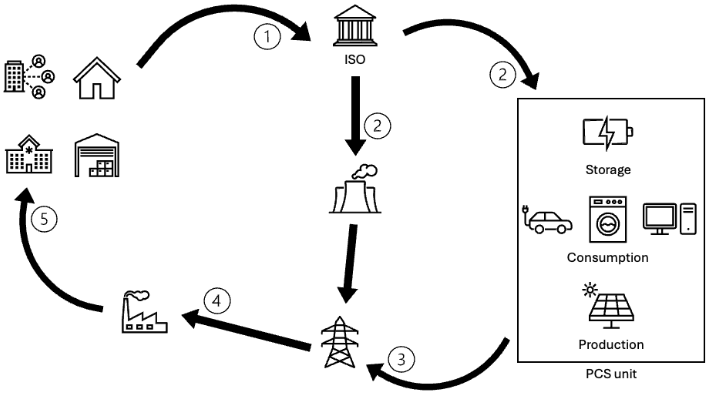

Figure 1: The day-ahead control cycle that repeats every 30-minutes: (1) ISO receives realized demand for the current time step. (2) ISO posts real-time buy/sell tariffs and issues dispatch directives to the controlled generators (3) GEAgents buy/sell power (4) If needed, peaker reserves are dispatched or curtailment is performed (5) Balanced power flows to consumers.

To address these challenges, the ISO adjusts electricity production plan, or *dispatch*, and feed-in and sell prices to influence independent market participants and align their behavior with grid operational objectives. Additionally, it retains access to reserves and peaking power plants, which can be activated to address unmet demand, ensuring both system stability and operational efficiency. The problem the ISO faces is thus one of cost optimization while satisfying the demand in the presence of strategic market players that aim to maximize their own profits. The scale and complexity of the problem make data-driven approaches, such as *reinforcement learning* (RL), especially suitable Perera & Kamalaruban (2021).

We make three key contributions. First, we build a ***multi-agent reinforcement learning*** **(MARL)** model that captures the incentives and rational decision-making of independent market participants. Leveraging these models, we then study the ISO's optimization problem under various assumptions, revealing how each setting shapes optimal dispatch and pricing policies. Finally, we offer a configurable, open-source grid simulator that supports diverse topologies and uncertainty patterns. Experiments across increasingly complex settings demonstrate that RL-driven agents can jointly optimise participant and ISO strategies, highlighting the promise of MARL for modern energy-market design.

2 Background and Related Work

Reinforcement Learning (RL) is a learning paradigm where an agent learns optimal behavior by interacting with an environment and receiving rewards or penalties for its actions Sutton & Barto (2018). Multi-agent reinforcement learning (MARL) extends RL to scenarios involving multiple autonomous agents that concurrently learn and make decisions within a shared or partially shared environment Albrecht et al. (2024). Each agent aims to maximize its own utility (typically measured as accumulated reward), but its actions can influence both its own outcomes and the outcomes of other agents, leading to complex emergent behaviors and the need for coordination and cooperation (see Appendix A for more detail).

The most common MARL model is the *stochastic game* (SG) (also known as emMarkov game or multi-agent MDP) Shapley (1953) defined as a tuple $\langle \mathcal{S}, \mathcal{A} = \{\mathcal{A}_i\}_{i=1}^{n}, \mathcal{T}, \mathcal{R} = \{\mathcal{R}_i\}_{i=1}^{n}, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the *joint action space* with $\mathcal{A}_i$ as the i^{th} agent action space s.t. $a \triangleq (a_1, a_2, \ldots, a_n)$ for $a \in \mathcal{A}$, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probability function $\mathcal{T}(s', a, s)$ such that $\forall s \in \mathcal{S}, \forall a \in \mathcal{A} : \sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s') = 1$, $\mathcal{R}$ is the *joint reward function* with $\mathcal{R}_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ as the i^{th} agent reward function, and $\gamma \in [0, 1)$ is the discount factor. A solution is a joint policy $\pi \triangleq (\pi_1, \ldots, \pi_n)$ associating each agent with policy $\pi_i : \mathcal{S} \times \mathcal{A}_i \rightarrow [0, 1]$ that specifies the probability of agent i taking an action at a given state. The joint policy should achieve certain conditions on the expected returns yielded to agents (e.g., Nash equilibrium) Albrecht et al. (2024). The value (utility) function $V_i^\pi(s)$ denotes the expected cumulative discounted reward agent i receives when starting in state s and the agents follow joint policy π thereafter. The action-value function or Q-value $Q_i^\pi(s, a)$ extends this notion by quantifying the expected value when performing a in s, and then continuing according to π. This general definition captures a variety of interactions and relationships that can exist between agents in collaborative, competitive, and mixed-incentive MARL settings.

MARL is particularly suitable for modeling energy systems and networks, since they are inherently multi-agent environments composed of diverse, distributed, and strategically autonomous entities, such as grid-edge components, utility companies, system operators, and market participants Zhu et al. (2023). These entities have different objectives, interact over shared physical and economic infrastructures, and must respond dynamically to system conditions, prices, and regulations. MARL provides a natural framework to model these interactions, enabling agents to learn adaptive policies, coordinate under uncertainty, and reason about both cooperative and competitive dynamics. Moreover, its ability to simulate emergent behavior and explore decentralized strategies makes it a powerful tool for both designing and analyzing modern energy systems.

Applications of RL and MARL in energy markets often assume a single, all-knowing controller optimizing the entire system. In such formulations, a central agent (analogous to an ISO) directly controls all generation and storage decisions using global information and perfect foresight, an assumption that is unattainable in practice. These centralized optimization models can yield system-level insights but cannot capture the strategic, profit-driven behavior of individual market participants Harder et al. (2023); Perera & Kamalaruban (2021). Moreover, as modern grids grow more heterogeneous and stochastic with high renewable penetration, a monolithic control scheme becomes impractical Wolgast & Nieße (2023). Recent studies emphasize that managing numerous distributed resources under uncertainty requires moving beyond one-size-fits-all control toward more decentralized decision-making structures Michailidis et al. (2025); Ahlqvist et al. (2022).

On the other end of the spectrum, many RL-based models use a fully decentralized approach in which each market participant (e.g. a storage unit owner or consumer) acts independently. In these formulations, multiple RL agents learn their own policies (for bidding, charging, discharging, etc.) based on price signals or local observations, without a central coordinator explicitly optimizing the whole systemWerner & Kumar (2023). This bottom-up approach reflects competitive markets by giving each market player its own profit-maximizing RL agentGuan et al. (2015); Vázquez-Canteli & Nagy (2019); Qiu et al. (2015). However, purely decentralized models typically assume the market rules or prices are exogenous or fixed Zhu et al. (2023); Ginzburg-Ganz et al. (2024); Perera & Kamalaruban (2021). In our model, the ISO acts as an active participant and directly shapes the market dynamics. Related efforts on dynamic dispatch and end-to-end RL in energy systems include Yang et al. (2021); Zhang et al. (2019), and comprehensive overviews of RL for power systems can be found in Ginzburg-Ganz et al. (2024).

From an algorithmic view the hard part is the *two-way* game: a learning ISO adjusts dispatch and the price pair ξ_t, ϕ_t each step, while strategic agents respond to maximise profit. Most work either treats the grid as one central optimiser or fixes ISO actions and lets agents learn in isolation; full bidirectional learning is rare Harder et al. (2023); Navon et al. (2024). Our framework closes that gap by explicitly modelling the feedback loop between an adaptive coordinator and autonomous market players, exactly the setting modern hybrid power markets require.

3 Energy Market Dynamics

Historically, the energy market comprised three principal components: power producers (e.g., power plants), power consumers (industrial and residential), and the ISO, responsible for market management and coordination. The producers typically used conventional coal-based generation and were either units under the full control of the ISO, or independent units that participated in the market but were regulated and bound by production agreements made for different temporal horizons.

In a typical *day-ahead market*, as depicted in Figure 1, the ISO predicts the following day's power demand (electricity consumption) and issues a *dispatch*, a production schedule, while considering operational constraints and generation costs. In addition to the generation of the predicted, or *nominal* demand, the ISO also manages the *reserve*, which sets a backup production capability for each time step. In real-time, the ISO is tasked with continuously maintaining a balance between demand and supply. If there is a surplus, energy is discharged, or *curtailed*. If production determined by the dispatch is not enough to cover the *realized demand*, reserves, which are more flexible but also more expensive and polluting, are deployed. Producers are then compensated based on the System Marginal Price (SMP) mechanism, calculated as the marginal cost of producing the final unit of energy required to satisfy system demand, based on the least-cost dispatch solution. In this work, we abstract the dispatch details and consider only the total amount and cost of power produced at each timestamp (see Appendix B and C for details on market dynamics and SMP computation, respectively).

Independent grid-edge GEAgents,private utilities and smart homes, now operate a single **Production–Consumption–Storage unit (PCS-unit)** that can generate (e.g. PV), consume, and store energy. Because they ignore dispatch orders and freely trade to maximise profit, the grid operator (ISO) can only shape their behaviour through prices. Its levers are the dispatch schedule Δ_t and the sell / feed-in tariffs ξ_t and ϕ_t set each interval t, chosen to balance supply and demand at minimum total cost. The sections that follow analyse this joint dispatch–pricing problem under progressively richer market assumptions.

In the deterministic setting, fully formulated in Appendix B , the ISO receives at the beginning of each episode the nominal production and reserve capabilities and costs for market participants, as well as the demand for all time steps in the horizon T. Based on this information and the operational constraints, it determines the scheduled Δ_t and prices $\xi_t(\cdot)$, $\phi_t(\cdot)$ for all timestamps $t \in [T]$ to minimize total costs. Formally,

$$\min C^{\text{total}} \;=\; \min\left[C^{\text{dispatch}} \;+\; \sum_{t=1}^{T} C_t^{\text{online}}\right] \qquad \text{(Deterministic ISO Objective)}$$

where C^{dispatch} is the total dispatch cost for the complete episode, and C_t^{online} is the online cost (including reserve cost) for time t.

Since all information is given in advance, the GEAgent can also compute its policy at the beginning of each episode and decide how much power to buy from (P_t^b), and sell to (P_t^s) the grid at every timestamp t to maximize its total revenue under its operational constraints. Formally:

$$\max \sum_{t=1}^{T} \Big(\phi_t\, P_t^s - \xi_t\, P_t^b\Big) \qquad \text{(Deterministic GEAgent Objective)}$$

In a stochastic extension of this setting, we account for the inability to exactly predict demand and production. In this case, it may be possible to estimate these distributions from historical data and observations using machine learning methods to improve decision-making under these forms of uncertainty. In this setting, fully formulated in Appendix B, the min and max objectives of the ISO and GEAgents are replaced by an expectation-based optimization.

Accounting for Strategic Demand: In modern energy systems, demand is not only stochastic but also strategic since GEAgents can intelligently manage the operation of devices and energy resources, in response to system-level signals. This *demand (load) flexibility* is reshaping energy markets by introducing new ways to contribute to their efficient and stable operation Charbonnier et al. (2022); Zhu et al. (2023). However, this shift also introduces challenges such as increased system complexity, uncertainty in demand forecasting, and the need for regulatory mechanisms to ensure fair and reliable participation.

In this extended setting, the ISO needs to determine the selling price ξ_t and feed-in prices ϕ_t for each t according to the demand D_t at time t while accounting for the GEAgents ability to sell, buy, and store power. From the perspective of the GEAgent, the price signals $\xi_t(P_t^s, P_t^b, \ldots)$ are exogenous signals set by the ISO , but they depend on the GEAgents' sales P_t^s and purchases P_t^b and other variables. This coupling results in a feedback mechanism where the player's actions influence the prices, and the prices, in turn affect the player's actions. This introduces a game-theoretic dimension where the GEAgents' decisions are influenced by the ISO 's pricing strategy and vice versa.

Formally, the GEAgent's input includes all the parameters that were relevant for the deterministic and stochastic settings, including the expected demand l_t and production g_t at time t. A key difference is that the selling price ξ_t and feed-in prices ϕ_t can be set either in advance or, depending on regulation, dynamically, in response to the market state. The objective of the GEAgent is now:

$$\max_{P_t^b, P_t^s} \mathbb{E}_{l_t, g_t} \left[\sum_{t=1}^{T} \left(\phi_t(P_t^s, P_t^b, \ldots) - \xi_t(P_t^s, P_t^b, \ldots) \right) \right] \qquad \text{(Strategic Player Objective)}$$

From the perspective of the ISO, as in the stochastic settings, it receives at the beginning of each episode (day) all the information about the GEAgents and the controlled producers and needs to determine the scheduled amount of production Δ_t for each timestamp. However, it is crucial to distinguish between two components of the demand. The **nominal demand** refers to the exogenous, inelastic portion of load that remains unaffected by local control strategies, real-time market incentives, or variations in renewable generation. In contrast **flexible demand**, refers to the portion of demand that can be adjusted in time, quantity, or pattern in response to external signals, such as price changes, grid conditions, or availability of renewable energy.

Since the ISO cannot loyally model the demand without considering the strategic nature of the GEAgents, optimization methods that are appropriate for deterministic and stochastic settings won't work here. Thus, as we specify in the next section, we model the market participants as RL agents.

4 The `Energy-Net` Simulator

In spite of a variety of simulators that currently exist Pigott et al. (2022); Moriyama (2018); Vázquez-Canteli et al. (2019); Marot (2021), there is no current framework that allows modeling the complex structure we want to account for and that is designed to work with off-the-shelf RLand MARL methods. We therefore develop a novel simulator, `Energy-Net`, that we will use to examine our proposed solutions. `Energy-Net` is a modular, discrete–time simulator of a hybrid electricity market. The environment we develop is flexible and adaptable, and can be used to accommodate different system configurations. At the core of the design of the software is a decoupling between the physical dynamics of the electrical system and the strategic agents, i.e., it is built around a strict *physics–agent split*. A high-fidelity physical core advances loads, renewables, batteries, and reserves, while the ISO and GEAgents interact only through a Gym-style `step()` interface. This design (i) lets us plug in any off-the-shelf RL algorithms without touching the power-system code, (ii) isolates market rules in a single controller module, and (iii) ensures that learned policies can affect the grid *only* via explicit levers, prices and dispatch tweaks, thus preserving physical realism while streamlining experimentation.

Building on the formal setting introduced in Appendix G, `Energy-Net` instantiates the 24-hour day-ahead electricity market. A single simulation episode therefore comprises T uniform intervals of length Δt (in our experiments T=48 and Δt = 30mins), together covering one 24-hour operational horizon. At each step $t \in \{1, \ldots, T\}$ the environment reveals the current forecast and grid state to the agents, applies their actions, propagates the physical dynamics, and returns next-state observations and rewards through the standard Gym `step` interface. See Appendix H for the full details.

5 Solution Approaches

The MARL formulation described in Appendix G provides an abstraction that captures the strategic, price-driven interactions that typify modern hybrid power systems. In this section, we present solution approaches that can be adopted by the market participants. Importantly, while our main challenge is in computing optimal market management approaches for the ISO, we must equip the GEAgents with the strongest policies to guarantee the ISO can predict their response to different price signals.

In principle, the deterministic and stochastic formulations described in Section 3 can be solved using state-space and dynamic programming methods, respectively (see Appendix D for an example formulation). Even if distributions are not fully known, it may be possible to learn them from data. Nevertheless, such methods are not appropriate to our problem, which is inherently challenging due to the agents' ability to strategically adapt their behavior and due to the dual-action learning structure, which operates across different time frames.

A specific challenge is that pricing may be dynamic and set at every time step, while the Δ_t action for each time step t is decided at the beginning of each episode. This temporal disparity adds a layer of complexity, as the reward for a Δ action is reflected only at the end of the episode. Moreover, determining Δ is a demanding task because it involves generating a time series output that must account for dynamic market conditions, which are influenced by behaviors of market participants. A further complication arises from the interdependence of these actions. Dispatch decisions are influenced by the market agents' responses to price signals, while optimal pricing strategies depend on real-time D_t and Δ_t outcomes.

Because the game is sequential (ISO first, GEAgent second) and highly non-linear, we iteratively train each of the policies with deep RL for continuous control in an online regime. If the agents' policies converge, it is toward a *practical* equilibrium in function-approximation space rather than a formal Nash point. in Section 6, we empirically examine this using our simulated environment described in the next section.

There are several abstractions that we can use to facilitate computation. One option is to make the problem easier is by abstracting away the dispatch optimization, which we denote as **dispatch abstraction**. In this simplified model the ISO has only control over the prices, and we assume that the ISO production Δ_t is fixed to be equal to the predicted demand $\hat{D}_t$.

Quadratic Pricing: We employ two pricing regimes, online dynamic and day-ahead tariffs. In settings restricted to day-ahead pricing, *quadratic pricing* allows the ISO to influence consumption and injection patterns through price curvature. Following Papadaskalopoulos & Strbac (2015), we impose a superlinear surcharge on purchases and a sublinear bonus on feed-in:

$$\xi_t = \alpha_0 + \alpha_1 \, P_t^b + \alpha_2 \, [P_t^b]^2, \quad \phi_t = \beta_0 + \beta_1 \, P_t^s + \beta_2 \, \sqrt{P_t^s},$$

where the six coefficients $(\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2)$ are fixed at the episode's outset for the subsequent T time steps. The superlinear term steepens the marginal purchase price, thereby discouraging demand spikes and reducing reliance on peaker reserves, while the sublinear feed-in adjustment tempers incentives for excessive injections, promoting smoother system operation (see Appendix E for full details and detailed examples).

6 Empirical Evaluation

The objective of our empirical evaluation is to assess the benefit of using our MARL formulation to optimize the policy of the ISO. For this, we use our `Energy-Net` environment to model and simulate the day-ahead electricity market[2].

[2] To respect the blind review process, our code base and complete results are in the supplementary material. All will be made public after acceptance.

Setup We evaluate our formulation from Appendix G and pricing schemes from Section 5 under a variety of scenarios. As discussed in Section 4, `Energy-Net` cleanly separates physical dynamics from agent logic. This allows us to stage the empirical study in three escalating phases of coordination for the ISO and GEAgents. First, in `ISO-Dispatch`, we trained and evaluated the ISO in isolation; all GEAgents were disabled, so the operator optimised its dispatch Δ_t under a stochastic yet *non-strategic* demand profile. Next, we enabled a PCS-unit[3] with a fixed, pre-defined charging trajectory and retrained the ISO, thereby quantifying the benefit of price coordination when storage is present but *non-adaptive*. We examined this setting with two pricing mechanisms: *online linear*, denoted `ISO-L`, and *quadratic*, denoted `ISO-Q`. We then allowed *both* agents to learn concurrently: the ISO tunes its real-time dispatch and tariffs, while the PCS-unit adapts its behavior to these market signals. In settings `Joint-Storage-L` and `Joint-Storage-Q` we examined the online and linear pricing, respectively, for a storage-only GEAgent, while in `Joint-PCS-L` and `Joint-PCS-Q`, we added production and consumption capabilities (see Appendix I for the full details of the setup). For each episode, we sample the *realized* demand from a Gaussian noise induced predicted demand for each time step t, and, when relevant, the realized load and production for the PCS-units. (see Table 3 in the appendix for a full description). We ran each training phase for 40 iterations with 4800 time steps each (1000 days) and was evaluated for 20 times. All settings were run using the same demand pattern and performance parameters described in Appendix I with Allocated resources of : 10 cores of Intel(R) Xeon(R) CPU E5-2683 v4 @ 2.10 GHz and 1 × NVIDIA GeForce GPU (12 GB). .

Results Due to space constraints, we present our full results in Appendix J and show here only our key findings. Our focus is on optimizing the ISO and measuring its ability to avoid failure and minimize cost, thus preferring to exploit renewable energy generated by the GEAgents and avoiding usage of reserves as much as possible. We therefore present in Table 4 the average energy usage achieved for all multi-agent settings compared to baseline `ISO-Dispatch`. To fully appreciate the effect of each agent setup, we present a breakdown of the total energy in MWh into three components: dispatch, reserve, and exchange (variance values in parentheses).

Results show that for settings `ISO-L` and `ISO-Q`, in which the GEAgent is fixed, the ISO manages to learn to exploit the power generated by the GEAgents instead of the reserves. In contrast, in `Joint-Storage-L` and `Joint-Storage-Q`, with a storage-only GEAgent the PCS-unit energy does not contribute to the overall efficiency. Instead, it increases the amount the ISO produces via dispatch to maintain stability. In Appendix J we show how this effect can be mitigated with different cost coefficients. Finally, for the complete setup of `Joint-PCS-L` and `Joint-PCS-Q`, where the GEAgents have consumption and production capabilities, we see a minimization of the reserve with quadratic pricing. To further demonstrate GEAgents contribution, Figure 7 depicts an episode from the `Joint-PCS-L` and `Joint-PCS-Q` settings. The difference between the dashed black line and the blue line (realized demand) represents the gap between the nominal predicted demand and the realized demand. The dispatch is represented by the light blue bars, while the total demand, including the flexible load of the GEAgents is depicted by the red line (total demand). As demonstrated in the figure, the reserve activation happens when the red line is *above* the dispatch bars, which is to be avoided. Overall, our experiments show that while fixed-generation players (`ISO-L`and `ISO-Q`) enable the ISO to substitute market output for reserves and storage-only players (`Joint-Storage-L`and `Joint-Storage-Q`) can unintentionally boost dispatch, it is only the combined consumption–production scenario (`Joint-PCS-L`) under a quadratic day-ahead tariff that suppresses reserve activation and maximizes system efficiency.

7 Conclusion

We demonstrate the benefit of modeling modern power systems MARL in which physical grid constraints, market signals, and heterogeneous agent behaviors interact in tightly coupled feedback

[3] Additional units can be added using the same interface; for clarity, we use one aggregated unit.

Table 1: Episode–total *energy* in MWh breakdown across scenarios.

Scenario	**Dispatch**	**Reserve**	**Exchange**
`ISO-Dispatch`	7 229.86 ± 38.29	249.41 ± 5.04	NA
`ISO-L`	7 282.34 ± 50.89	176.05 ± 19.07	800 ± 0
`ISO-Q`	7 506.98 ± 35.02	121.07 ± 3.78	800 ± 0
`Joint-Storage-L`	8 126.13 ± 1.07	148 ± 0.94	0 ± 0
`Joint-Storage-Q`	8 126.21 ± 1.01	148 ± 1.06	0 ± 0
`Joint-PCS-L`	7 322.44 ± 36.02	168.47 ± 4.14	442.14 ± 9.61
`Joint-PCS-Q`	7 450.62 ± 36.43	**117 ± 2.04**	324 ± 8.40

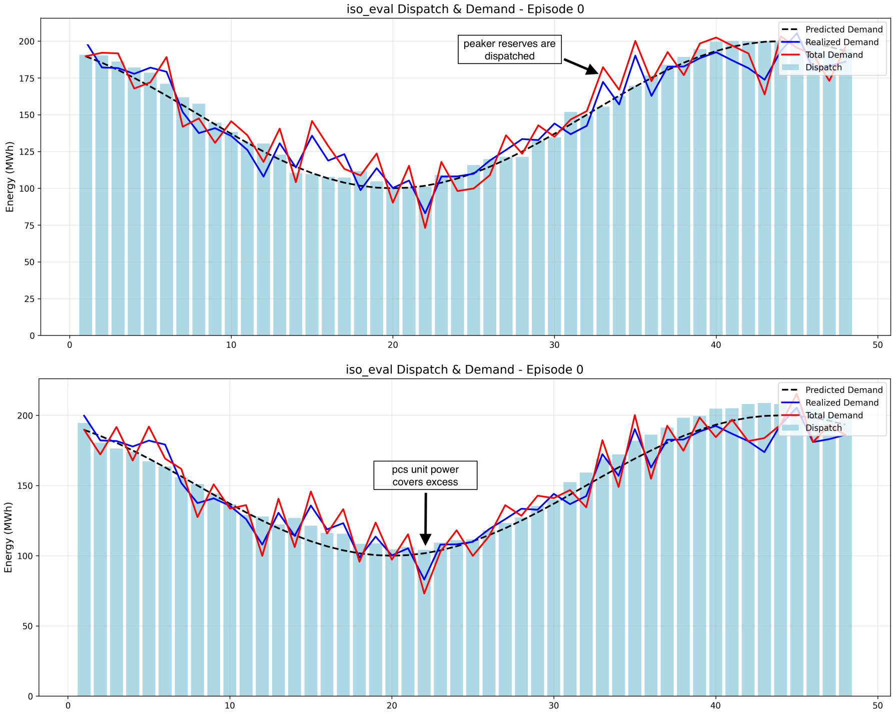

Figure 2: Episode-level dispatch and realized demand under scenario `Joint-PCS-L` (online linear pricing) at the top, and scenario `Joint-PCS-Q` (quadratic pricing) at the bottom.

loops. We design our framework to capture both nominal and flexible demand, and enable realistic and robust evaluation of decentralized control strategies and pricing mechanisms using a new simulation environment we developed. Our results show that strategically coordinated ISO policies working with price-responsive grid-edge agents can reduce reserve requirements and carbon intensity.

Together with these achievements, our experiments reveal the fragility of current deep-RL policies: modest forecasting errors can lead to supply shortfalls or excessive generation. Addressing this brittleness remains a key research priority. Another challenge lies in scaling the approach operational grids. This will require hierarchical or federated MARL architectures and hardware-in-the-loop testing. Finally, while algorithmic coordination can reduce reserve usage and lower tariffs, distribution benefits are unlikely to be uniform. Ensuring fairness and transparency is a challenge that will need to be addressed.

Appendix

A RL and MARL

A Reinforcement Learning (RL) problem can be defined as a Markov Decision Process (MDP) represented by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, where:

- $\mathcal{S}$ is the set of states,
- $\mathcal{A}$ is the set of actions,
- $\mathcal{P}(s' \mid s, a)$ is the transition probability from state s to s' under action a,
- $\mathcal{R}(s, a)$ is the reward function,
- $\gamma \in [0, 1]$ is the discount factor.

The goal is to find a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the expected cumulative reward:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R_t \right],$$

where R_t is the reward received at time step t. It is assumed that the MDP is too large to efficiently compute π^*, so approximation methods are employed to estimate it. These methods often involve learning value functions or directly optimizing parameterized policies using sampled interactions with the environment.

The problem can be modeled as a Markov Decision Process (MDP), defined by the tuple:

$$\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$$

where:

- $\mathcal{S}$: The set of states, defined by $\mathcal{S} = \{(t, \sigma_t) \mid t = 1, \ldots, T,\ 0 \leq \sigma_t \leq S_{\max}\}$,
- $\mathcal{A}$: The set of actions, where each action is represented by the pair (P_t^b, P_t^s),
- $\mathcal{P}(s' \mid s, a)$: The state transition function, given by:

$$\mathcal{P}(s' \mid s, a) = \Pr(\sigma_{t+1} \mid \sigma_t, P_t^b, P_t^s),$$

- $\mathcal{R}(s, a)$: The reward function:

$$\mathcal{R}(s, a) = \phi_t(P_t^s) - \xi_t(P_t^b),$$

- γ: The discount factor, $\gamma \in [0, 1]$, which determines the relative importance of future rewards.

The goal is to find an optimal policy π^* that maximizes the expected cumulative reward:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=1}^{T} \gamma^{t-1} \mathcal{R}(s_t, a_t) \right],$$

where:

- $s_t = (t, \sigma_t)$ is the state at time t,
- $a_t = (P_t^b, P_t^s)$ is the action at time t,
- $\mathcal{R}(s_t, a_t)$ is the immediate reward obtained from taking action a_t in state s_t.

rl and marl algorithms can be broadly categorized as model-free, which learn policies directly from experience without modeling the environment, and model-based, which learn or use environment models to plan or simulate outcomes. Model-free methods (e.g., value-based or policy gradient) tend to be more scalable but sample-inefficient, while model-based methods improve sample efficiency and enable planning but struggle with modeling complex dynamics Albrecht et al. (2024).

Reinforcement Learning (rl) is a learning paradigm where an agent learns optimal behavior by interacting with an environment and receiving rewards or penalties for its actions Sutton & Barto (2018).

Multi-agent RL (marl) extends rl to scenarios involving multiple autonomous agents that concurrently learn and make decisions within a shared or partially shared environment. Each agent aims to maximize its own utility (typically measured as accumulated reward), but its actions can influence both its own outcomes and the outcomes of other agents, leading to complex emergent behaviors and the need for coordination and cooperation.

marl is particularly suitable for modeling energy systems and networks, since they are inherently multi-agent environments composed of diverse, distributed, and strategically autonomous entities, such as grid-edge components and prosumers, utility companies, system operators, and market participants. These entities have different objectives, interact over shared physical and economic infrastructures, and must respond dynamically to system conditions, prices, and regulations. MARL provides a natural framework to model these interactions, enabling agents to learn adaptive policies, coordinate under uncertainty, and reason about both cooperative and competitive dynamics. Moreover, its ability to simulate emergent behavior and explore decentralized strategies makes it a powerful tool for both designing and analyzing modern energy systems.

The most common marl model is the Stochastic Game (also known as Markov Game or Multi-agent MDP) Shapley (1953) defined as a tuple $\langle \mathcal{S}, \mathcal{A} = \{\mathcal{A}_i\}_{i=1}^{n}, \mathcal{T}, \mathcal{R} = \{\mathcal{R}_i\}_{i=1}^{n}, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the *joint action space* with $\mathcal{A}_i$ as the i^{th} agent action space s.t. $a \triangleq (a_1, a_2, \ldots, a_n)$ for $a \in \mathcal{A}$, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is the transition probability function $\mathcal{T}(s', a, s)$ such that $\forall s \in \mathcal{S}, \forall a \in \mathcal{A} : \sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s') = 1$, $\mathcal{R}$ is the *joint reward function* with $\mathcal{R}_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ as the i^{th} agent reward function, and $\gamma \in [0, 1)$ is the discount factor. A solution is a joint policy $\pi \triangleq (\pi_1, \ldots, \pi_n)$ associating each agent with policy $\pi_i : \mathcal{S} \times \mathcal{A}_i \to [0, 1]$ that specifies the probability of agent i taking an action at a given state. The joint policy should achieve certain conditions on the expected returns yielded to agents (e.g., Nash equilibrium) Albrecht et al. (2024). The value (utility) function $V_i^{\pi}(s)$ denotes the expected cumulative discounted reward agent i receives when starting in state s and the agents follow joint policy π thereafter. The action-value function or Q-value $Q_i^{\pi}(s, a)$ extends this notion by quantifying the expected value when performing a in s, and then continuing according to π. A Multi-agent Partially Observed MDP (or Partially Observable Stochastic Game) also includes for each agent observation set O_i and a sensor function $\mathcal{O}_i : \mathcal{A} \times \mathcal{S} \times O_i \to [0, 1]$.

This general definition captures a variety of interactions and relationships that can exist between agents in collaborative, competitive, and mixed-incentive MARL settings. Complex agent interactions may give rise to behaviors that are difficult to anticipate by simply examining each agent in isolation. Thus, despite the potential to solve complex problems across various domains, marl faces various significant challenges that stem from aspects such as scale, conflicting goals of self-interested agents, and the concurrent learning of the different agents Albrecht et al. (2024). All these are relevant to MARL in general but are particularly relevant to energy networks with the added need to account for the dynamics of the physical environment and the effect decisions may have on the functioning of the electricity network.

RL and MARL algorithms can be broadly categorized as model-free, which learn policies directly from experience without modeling the environment, and model-based, which learn or use environment models to plan or simulate outcomes. Model-free methods (e.g., value-based or policy gradient) tend to be more scalable but sample-inefficient, while model-based methods improve sample efficiency and enable planning but struggle with modeling complex dynamics.

B Energy Market Dynamics

B.1 Energy Markets and the Dispatch Problem

Historically, the energy market comprised three principal components: power producers (e.g., power plants), power consumers (industrial and residential), and the ISO, responsible for market management and coordination. The producers typically used conventional coal-based generation and were either units under the full control of the ISO, or independent units that participated in the market but that were fully regulated, i.e., bound by production agreements made with the .

A typical structure of a market was based on the *day-ahead market* in which the ISOpredicts the following day's power demand and issues a *dispatch*, an offline production schedule to each producer while considering operational constraints and generation costs. The dispatch traditionally divides the 24-hour planning horizon into 48 discrete half-hour time periods. In addition to the generation of the predicted, or *nominal* demand, the ISO also manages the *reserve*, which sets a backup production capability for each time step. If in real-time the controlled production determined by the dispatch is not enough to cover the realized demand, reserves, which are more flexible but also more expensive and polluting, are activated by an online controller. Producers are then compensated based on the System Marginal Price (SMP) mechanism, which is calculated as the marginal cost of producing the final unit of energy required to satisfy system demand, based on the least-cost dispatch solution (See Appendix C). For the purposes of this work we abstract the dispatch details, and consider only the total amount of power produced at each timestamp, as well as its total cost to the ISO with no regard to the inner structure of the dispatch.

Recent reforms in the power market have introduced independent grid-edge market players, which we denote as **GEAgent**s, including private electric companies and smart homes. These new market players possess the ability to produce electricity, manage internal consumption, and utilize power storage capabilities. Unlike traditional controlled producers, they are not legally required to adhere to dispatch instructions and may buy from or sell to the grid at will to maximizing their profits. We assume GEAgents are rational, so the natural way for the ISO to induce desired behaviors of the market players is via price signals. In real-time operations, the ISO manages the grid by buying electricity from power producers and selling it to consumers. The selling price at time t, denoted as ξ_t, and the feed-in price, denoted as ϕ_t, are the primary tools for market control.

The GEAgent models are essential for the ISO's planning, as they capture participant strategies and behaviors that influence the grid's supply-demand balance. These models enable the ISO to design pricing mechanisms, such as sell prices and feed-in tariffs, to align player incentives with grid stability and efficiency. We classify market player behaviors in increasingly realistic environments, starting with simpler cases to build intuition before progressing to more complex scenarios, as the problems share similar structures. In correspondence with current energy markets, each GEAgent operates a Production-Consumption-Storage (PCS) unit, which can produce (e.g., via pv), consume (e.g., via electrical appliances) and store (e.g., via a battery) energy. It aims at maximizing its profit over the period in question.

To determine dispatch and pricing, the ISO utilizes demand predictions for the subsequent 24 hours, denoted $\hat{D}_t$, where t represents the time interval. Based on these predictions, the ISO determines a scheduled production dispatch Δ_t for each timestamp. It also determines for each time step how much reserve to guarantee, specifying the standby capacity to maintain in response to unexpected demand surges or generation outages. Reserve energy enhances grid reliability but can be highly polluting when supplied by fossil-fuel generators, which operate inefficiently and emit more greenhouse gases.

In real-time operations, the ISO manages the grid by buying electricity from power producers and selling it to consumers. The selling price at time t, denoted as ξ_t, and the feed-in price, denoted as ϕ_t, are the primary tools for market control.

The electricity market includes n independent agents representing the GEAgents, indexed by $i \in \{1, \ldots, n\}$, who operate autonomously to maximize their profits. The ISO has no direct control over these agents, and their interactions are governed by market dynamics, which are influenced by various regulations. These regulations, coupled with non-economic factors, significantly shape the cost structure of the system. However, the ISO can compute costs based on relevant inputs and adapt its computational models dynamically to reflect changes in regulations or legislation.

In this work, we suggest using RL-to model the market participants and ways for the ISO to control the dispatch Δ and price signals ξ, ϕ to minimize the total costs for the ISO (thus the taxpayers) while satisfying the demand. A key challenge is that this needs to be done while taking market players' strategic behavior into account.

To support optimizing the ISO's behavior, we analyze how market players react to prices in increasingly complex settings, from deterministic to stochastic and strategic environments.

B.2 Market Participants

The GEAgent models are essential for the ISO's planning, as they capture participant strategies and behaviors that influence the grid's supply-demand balance. These models enable the ISO to design pricing mechanisms, such as sell prices and feed-in tariffs, to align player incentives with grid stability and efficiency. We classify market player behaviors in increasingly realistic environments, starting with simpler cases to build intuition before progressing to more complex scenarios, as the problems share similar structures. In correspondence with current energy markets, each GEAgent operates a Production-Consumption-Storage (PCS) unit, which can produce (e.g., via PV), consume (e.g., via electrical appliances) and store (e.g., via a battery) energy. It aims at maximizing its profit over the period in question.

Having settled on market players', we proceed to present the task that the ISO faces. The ISO is tasked with meeting electricity demand at all times. To achieve this, the ISO controls the dispatch of electricity generation. While the specifics of which power plant generates how much power are abstracted, the total scheduled electricity production is determined for each time step, ensuring sufficient supply to meet anticipated demand.

The ISO aims to maximize its utility, which may include balancing grid supply and demand, minimizing operational costs, or promoting renewable integration.

The total cost incurred by the ISO increases marginally due to the characteristics of the SMP mechanism. The SMP prioritizes electricity from the cheapest sources first, resulting in higher costs for additional megawatts of production as cheaper resources are exhausted. Additionally, sharp changes in production across time steps introduce significant costs due to ramp-up and cool-down constraints of power plants. These transitions strain generation units, necessitating increased operational expenses. The ISO incorporates these costs into pricing to discourage abrupt fluctuations, maintaining grid stability.

To influence the behavior of market players, the ISO offers sell prices and feed-in tariffs. These prices act as economic signals, encouraging players to adjust their electricity consumption, production, and storage behaviors in alignment with grid stability and efficiency goals. By strategically setting these prices, the ISO aims to optimize the overall operation of the electricity market under a hybrid public-private model.

This is no longer true: In what follows, we examine three dimensions of complexity: (1) the nature of demand, encompassing three levels—deterministic and known, stochastic, and strategic, (2) the decision types, including buy/sell and dispatch, and (3) the decision horizon; are decisions made offline (for the entire 24-hour horizon), or online (e.g., every 30 minutes). what are the decision horizons we consider what are the decision horizons we consider

In what follows, we examine three levels of complexity that are associated with the nature and pattern of the demand (consumption): deterministic and known, stochastic, and strategic.

B.3 Deterministic Setting

As a first step, we consider a a fully deterministic environment, where the demand is fully known in advance and the prices are set in advance (at time 0 of every day).

- Storage capacity: $S_{\max}$.
- Maximum charging rate: $C_{\max}$.
- Maximum discharging rate: $D_{\max}$.
- Initial storage state of charge: σ_0.

- Selling price levels ξ_t set by the ISOfor each time interval and known in advance to the player.
- Feed-in prices ϕ_t set by the ISOand known in advance to the player as well.

Since all information is given in advance, the GEAgent can compute optimal policies at time-step 0. A GEAgent must decide how much power to buy from (P_t^b), and sell to (P_t^s) the grid at every timestamp t to maximize its total revenue under its operational constraints. Formally:

$$\max \sum_{t=1}^{T} \left(\phi_t(P_t^s) - \xi_t(P_t^b)\right) \tag{Deterministic GEAgent Objective}$$

Subject to:

1. **Power Balance Constraints:**
   At each time t, the power bought or sold must meet the demand, including charging:

$$\forall t : P_t^b - P_t^s = l_t + (\sigma_t - \sigma_{t-1}) \tag{C1}$$

   Here we assume a lossless battery.
2. **Storage Capacity Constraints:**
   The storage level must remain within capacity limits:

$$\forall t : 0 \leq \sigma_t \leq S_{\max} \tag{C2}$$

3. **Charging and Discharging Rate Constraints:**

$$\forall t : -D_{\max} \leq P_t^b - (l_t + P_t^s) \leq C_{\max} \tag{C3}$$

4. **Non-Negativity Constraints:**

$$\forall t : P_t^b, P_t^s, \sigma_t \geq 0 \tag{C4}$$

5. **No Simultaneous Charging and Discharging:**

$$\forall t : P_t^b \cdot P_t^s = 0 \tag{C5}$$

distinguish here between the producers and the players distinguish between fixed parameters and inputs

ISO In the deterministic case, at the start of the planning horizon (timestamp 0), the ISO receives the following inputs:

- Nominal? Demand D_t for all timestamps in the horizon.
- Reserve activation cost C_{reserve}.
- The number of GEAgents N participating in the market.
- The maximum discharge rates of each market player $D^i_{\max}$.

Based on this information, the ISO determines the scheduled amount of production Δ_t and prices $\xi_t(\cdot)$, $\phi_t(\cdot)$ for all timestamps $t \in [T]$ ahead. Then, at each timestamp t market players can respond to the prices by buying or selling power to the grid, contributing a net power demand P_t^{net}. If the net demand after accounting for P_t^{net} exceeds the scheduled production Δ_t, the ISO activates reserves or peaker plants to cover the shortfall. If the market players are assumed to be rational, and the ISOmakes the prices public at $t = 0$, the market players are solving the deterministic problem as presented in Section B.3, and the ISO can run the simulation of the market players to optimize the dispatch and the price signal.

The ISO aims to minimize its total costs,

$$\min C^{\text{total}} = \min \left[C^{\text{dispatch}} + \sum_{t=1}^{T} C_t^{\text{online}} \right] \qquad \text{(ISO objective)}$$

where:

- **Cost of the Dispatch Schedule ($C^{\textbf{dispatch}}$)**:

$$C^{\text{dispatch}} = \sum_{t=1}^{T} C(\Delta_t) + \sum_{t=2}^{T} \rho(\Delta_0, \ldots, \Delta_t),$$

where ρ is a penalty function that can be tailored to various performance criteria, e.g., for penalizing sharp changes in dispatch levels between consecutive periods.

- **Online Cost per Timeframe ($C_t^{\textbf{online}}$)**: The sum of the market cost and the reserve activation cost:

$$C_t^{\text{online}} = C_t^{\text{market}} + C_t^{\text{reserve}}(\max(0, D_t - P_t^{\text{net}} - \Delta_t)),$$

Notably, we assume that all demand must be met, a constraint that can be relaxed if needed.

- **Market Cost per Timeframe ($C_t^{\textbf{market}}$)**: Payments to market players for the power they sell to the grid net of the revenue from selling the power to market players:

$$C_t^{\text{market}} = \sum_i \phi_t^{(i)}(s_t^{(i)}) - \sum_i \xi_t^{(i)}(b_t^{(i)})$$

where $\phi_t^{(i)}$ is the feed-in tariff offered to player i at time t, and $s_t^{(i)}$ is the amount of power sold by player i to the grid.

Note that this problem is unconstrained, since we assume that when the demand is not met by the production and the market, the ISO operates the reserves. The incentive to meet the demand using nominal generation is encapsulated ? in the typically high costs associated with activating the reserves.

B.4 Accounting for Stochasticity

Real-world systems are inherently stochastic, requiring models to account for uncertainty. Key sources of randomness include:

- Internal load variability,
- Renewable production fluctuations,
- Price changes driven by external demand uncertainty.

All these may lead to an inability to exactly predict the demand that will be needed.

From the point of view of the GEAgent, the main source of uncertainty can come from its To address this, the objective function is reformulated as:

$$\max \mathbb{E}_{l_t, \xi_t} \left[\sum_{t=1}^{T} \left(\phi_t(P_t^s) - \xi_t(P_t^b) \right) \right]. \qquad \text{(Stochastic Player Objective)}$$

At each timestamp t, the player observes the realizations of l_t, g_t, and ξ_t before deciding on P_t^b, and P_t^s.

In a stochastic environment, the distributions of l_t, g_t, and ξ_t may be unknown. If this is the case, the player can estimate these distributions from historical data and observations using machine learning methods to improve decision-making under these forms of uncertainty.

Figure out how we can deal with stochastic environments from the side of the ISO The main change becomes the uncertainty about the demand

In this case, we have two options, depending on when decisions need to be made.

B.5 Accounting for Load Flxibility and Strategic Demand

So far, we considered settings in which all participants were aiming to maximize their revenue (and minimize cost) while considering the deterministic or stochastic information that is received at time step 0, i.e., at the beginning of the daily episode. This meant that prices and dispatch decisions are made at the start of each episode, with the real-time decisions limited to reserve activation or curtailment (energy discharge) actions in response to unpredictable demand and the requirement to maintain stability.

In modern energy systems, demand is not only stochastic but also strategic. This is because grid-edge agents can intelligently manage the operation of devices and distributed energy resources (DERs), in response to system-level signals, such as prices, frequency, or voltage. This *load flexibility* is reshaping energy markets by introducing new ways by which grid-edge agents can contribute to the efficient and stable operation of the network Charbonnier et al. (2022); Zhu et al. (2023). However, this shift also introduces challenges such as increased system complexity, uncertainty in demand forecasting, and the need for regulatory mechanisms to ensure fair and reliable participation.

In this extended setting, the ISO aims to maximize its utility, but needs to determine the selling price ξ_t and feed-in prices ϕ_t for each time step t according to the demand D_t at time t. The key challenge is that D_t now includes the GEAgents ability to sell, buy, and store power. From the perspective of the GEAgent, the price signals $\xi_t(P_t^s, P_t^b, \ldots)$ represent the exogenous prices set by the ISO , which depend on the player's sales P_t^s and purchases P_t^b as well as other variables. This coupling results in a feedback mechanism where the player's actions influence the prices, and the prices in turn affect the player's actions. This introduces a game-theoretic dimension to the problem that the market player faces, where the player's decisions on P_t^b, and P_t^s are influenced by the ISO's pricing strategy and vice versa.

It is important to clarify what the possibilities are that are available to the ISO with regard to the dispatch and pricing decisions it can make. This is not only a technical question, but a regulatory and policy-making question that needs to be accounted for. Two common approaches are day-ahead and dynamic pricing.

Formally, the GEAgent's input includes timesteps $t = 1, 2, \ldots, T$ GEAgent's load: l_t, storage capacity: $S_{\max}$, maximum charging rate: $C_{\max}$, maximum discharging rate: $D_{\max}$, current storage state of charge: σ_0 as defined in sections B.3 and B.4. The key difference is that now the selling price ξ_t and feed-in prices ϕ_t can be set by ISOin advance or in a dynamic way, in response to the market state.

The objective is now:

$$\max_{P_t^b, P_t^s} \mathbb{E}_{l_t, g_t} \left[\sum_{t=1}^{T} \left(\phi_t(P_t^s, P_t^b, \ldots) - \xi_t(P_t^s, P_t^b, \ldots) \right) \right]. \qquad \text{(Strategic Player Objective)}$$

The ISO at the start of the planning horizon (timestamp 0), the ISOreceives the following inputs:

- The cost function of the production $C(\Delta_t)$ for each t.
- Predicted demand $\hat{D}_t$ for all timestamps in the horizon.
- Reserve activation cost per unit C_{reserve}.
- The number of market players N participating in the market.
- The maximum discharge rates of each market player $D_{\max}^i$.

Based on this information, the ISO determines the scheduled amount of production Δ_t for each timestamp in the horizon. Here, it is crucial to distinguish between nominal and flexible demand components. Nominal demand, denoted D refers to the exogenous, inelastic portion of load at each grid node that remains unaffected by local control strategies, real-time market incentives, or variations in renewable generation. In contrast, *flexible demand*, denoted l, refers to the portion of demand (electricity consumption) that can be adjusted in time, quantity, or pattern in response to external signals—such as price changes, grid conditions, or availability of renewable energy.

The objective of the ISO now becomes

$$\min \mathbb{E}_{D,l}\left[C^{\text{dispatch}} + \sum_{t=1}^{T} C_t^{\text{online}}\right] \tag{O2}$$

Since it is impossible for the ISO to precisely model market players' demand without considering its strategic nature, optimization methods that are appropriate for deterministic and stochastic settings won't work here. Thus, as we specify in the next section, we model the market using RL.

B.6 Deterministic Setting

As a first step, we consider a a fully deterministic environment, where the demand is fully known in advance and the prices are set in advance (at time 0 of every day).

- Storage capacity: $S_{\max}$.
- Maximum charging rate: $C_{\max}$.
- Maximum discharging rate: $D_{\max}$.
- Initial storage state of charge: σ_0.
- Selling price levels ξ_t set by the ISOfor each time interval and known in advance to the player.
- Feed-in prices ϕ_t set by the ISOand known in advance to the player as well.

Since all information is given in advance, the GEAgent can compute optimal policies at time-step 0. A GEAgent must decide how much power to buy from (P_t^b), and sell to (P_t^s) the grid at every timestamp t to maximize its total revenue under its operational constraints. Formally:

$$\max \sum_{t=1}^{T} \left(\phi_t(P_t^s) - \xi_t(P_t^b)\right) \tag{Deterministic GEAgent Objective}$$

Subject to:

1. **Power Balance Constraints:**
   At each time t, the power bought or sold must meet the demand, including charging:

$$\forall t : P_t^b - P_t^s = l_t + (\sigma_t - \sigma_{t-1}) \tag{C1}$$

   Here we assume a lossless battery.
2. **Storage Capacity Constraints:**
   The storage level must remain within capacity limits:

$$\forall t : 0 \leq \sigma_t \leq S_{\max} \tag{C2}$$

3. **Charging and Discharging Rate Constraints:**

$$\forall t : -D_{\max} \leq P_t^b - (l_t + P_t^s) \leq C_{\max} \tag{C3}$$

4. **Non-Negativity Constraints:**

$$\forall t : P_t^b, P_t^s, \sigma_t \geq 0 \tag{C4}$$

5. **No Simultaneous Charging and Discharging:**

$$\forall t : P_t^b \cdot P_t^s = 0 \tag{C5}$$

distinguish here between the producers and the players distinguish between fixed parameters and inputs

ISO In the deterministic case, at the start of the planning horizon (timestamp 0), the ISOreceives the following inputs:

- Nominal? Demand D_t for all timestamps in the horizon.
- Reserve activation cost C_{reserve}.
- The number of GEAgents N participating in the market.
- The maximum discharge rates of each market player $D^i_{\max}$.

Based on this information, the ISO determines the scheduled amount of production Δ_t and prices $\xi_t(\cdot)$, $\phi_t(\cdot)$ for all timestamps $t \in [T]$ ahead. Then, at each timestamp t market players can respond to the prices by buying or selling power to the grid, contributing a net power demand P_t^{net}. If the net demand after accounting for P_t^{net} exceeds the scheduled production Δ_t, the ISO activates reserves or peaker plants to cover the shortfall. If the market players are assumed to be rational, and the ISO makes the prices public at $t = 0$, the market players are solving the deterministic problem as presented in Section B.3, and the ISO can run the simulation of the market players to optimize the dispatch and the price signal.

The ISO aims to minimize its total costs,

$$\min C^{\text{total}} = \min \left[C^{\text{dispatch}} + \sum_{t=1}^{T} C_t^{\text{online}} \right] \tag{ISO objective}$$

where:

- **Cost of the Dispatch Schedule (C^{dispatch}):**

  $$C^{\text{dispatch}} = \sum_{t=1}^{T} C(\Delta_t) + \sum_{t=2}^{T} \rho(\Delta_0, \ldots, \Delta_t),$$

  where ρ is a penalty function that can be tailored to various performance criteria, e.g., for penalizing sharp changes in dispatch levels between consecutive periods.
- **Online Cost per Timeframe (C_t^{online}):** The sum of the market cost and the reserve activation cost:

  $$C_t^{\text{online}} = C_t^{\text{market}} + C_t^{\text{reserve}}(\max(0, D_t - P_t^{\text{net}} - \Delta_t)),$$

  Notably, we assume that all demand must be met, a constraint that can be relaxed if needed.
- **Market Cost per Timeframe (C_t^{market}):** Payments to market players for the power they sell to the grid net of the revenue from selling the power to market players:

  $$C_t^{\text{market}} = \sum_i \phi_t^{(i)}(s_t^{(i)}) - \sum_i \xi_t^{(i)}(b_t^{(i)})$$

  where $\phi_t^{(i)}$ is the feed-in tariff offered to player i at time t, and $s_t^{(i)}$ is the amount of power sold by player i to the grid.

Note that this problem is unconstrained, since we assume that when the demand is not met by the production and the market, the ISO operates the reserves. The incentive to meet the demand using nominal generation is encapsulated ? in the typically high costs associated with activating the reserves.

B.7 Accounting for Stochasticity

Real-world systems are inherently stochastic, requiring models to account for uncertainty. Key sources of randomness include:

- Internal load variability,
- Renewable production fluctuations,
- Price changes driven by external demand uncertainty.

All these may lead to an inability to exactly predict the demand that will be needed.

From the point of view of the GEAgent, the main source of uncertainty can come from its To address this, the objective function is reformulated as:

$$\max \mathbb{E}_{l_t,\xi_t} \left[\sum_{t=1}^{T} \left(\phi_t(P_t^s) - \xi_t(P_t^b) \right) \right]. \qquad \text{(Stochastic Player Objective)}$$

At each timestamp t, the player observes the realizations of l_t, g_t, and ξ_t before deciding on P_t^b, and P_t^s.

In a stochastic environment, the distributions of l_t, g_t, and ξ_t may be unknown. If this is the case, the player can estimate these distributions from historical data and observations using machine learning methods to improve decision-making under these forms of uncertainty.

Figure out how we can deal with stochastic environments from the side of the ISO The main change becomes the uncertainty about the demand

In this case, we have two options, depending on when decisions need to be made.

B.8 Accounting for Load Flxibility and Strategic Demand

So far, we considered settings in which all participants were aiming to maximize their revenue (and minimize cost) while considering the deterministic or stochastic information that is received at time step 0, i.e., at the beginning of the daily episode. This meant that prices and dispatch decisions are made at the start of each episode, with the real-time decisions limited to reserve activation or curtailment (energy discharge) actions in response to unpredictable demand and the requirement to maintain stability.

In modern energy systems, demand is not only stochastic but also strategic. This is because grid-edge agents can intelligently manage the operation of devices and distributed energy resources (DERs), in response to system-level signals, such as prices, frequency, or voltage. This *load flexibility* is reshaping energy markets by introducing new ways by which grid-edge agents can contribute to the efficient and stable operation of the network Charbonnier et al. (2022); Zhu et al. (2023). However, this shift also introduces challenges such as increased system complexity, uncertainty in demand forecasting, and the need for regulatory mechanisms to ensure fair and reliable participation.

In this extended setting, the ISO aims to maximize its utility, but needs to determine the selling price ξ_t and feed-in prices ϕ_t for each time step t according to the demand D_t at time t. The key challenge is that D_t now includes the GEAgents ability to sell, buy, and store power. From the perspective of the GEAgent, the price signals $\xi_t(P_t^s, P_t^b, \ldots)$ represent the exogenous prices set by the ISO , which depend on the player's sales P_t^s and purchases P_t^b as well as other variables. This coupling results in a feedback mechanism where the player's actions influence the prices, and the prices in turn affect the player's actions. This introduces a game-theoretic dimension to the problem that the market player faces, where the player's decisions on P_t^b, and P_t^s are influenced by the GSO's pricing strategy and vice versa.

It is important to clarify what the possibilities are that are available to the ISO with regard to the dispatch and pricing decisions it can make. This is not only a technical question, but a regulatory and policy-making question that needs to be accounted for. Two common approaches are day-ahead and dynamic pricing.

Formally, the GEAgent's input includes timesteps $t = 1, 2, \ldots, T$ GEAgent's load: l_t, storage capacity: $S_{\max}$, maximum charging rate: $C_{\max}$, maximum discharging rate: $D_{\max}$, current storage state of charge: σ_0 as defined in sections B.3 and B.4. The key difference is that now the selling price ξ_t and feed-in prices ϕ_t can be set by ISOin advance or in a dynamic way, in response to the market state.

The objective is now:

$$\max_{P_t^b, P_t^s} \mathbb{E}_{l_t, g_t} \left[\sum_{t=1}^{T} \left(\phi_t(P_t^s, P_t^b, \ldots) - \xi_t(P_t^s, P_t^b, \ldots) \right) \right]. \qquad \text{(Strategic Player Objective)}$$

The ISO at the start of the planning horizon (timestamp 0), the ISOreceives the following inputs:

- The cost function of the production $C(\Delta_t)$ for each t.
- Predicted demand $\hat{D}_t$ for all timestamps in the horizon.
- Reserve activation cost per unit C_{reserve}.
- The number of market players N participating in the market.
- The maximum discharge rates of each market player $D_{\max}^i$.

Based on this information, the ISO determines the scheduled amount of production Δ_t for each timestamp in the horizon. Here, it is crucial to distinguish between nominal and flexible demand components. Nominal demand, denoted D refers to the exogenous, inelastic portion of load at each grid node that remains unaffected by local control strategies, real-time market incentives, or variations in renewable generation. In contrast, *flexible demand*, denoted l, refers to the portion of demand (electricity consumption) that can be adjusted in time, quantity, or pattern in response to external signals—such as price changes, grid conditions, or availability of renewable energy.

The objective of the ISO now becomes

$$\min \mathbb{E}_{D,l} \left[C^{\text{dispatch}} + \sum_{t=1}^{T} C_t^{\text{online}} \right] \qquad \text{(Stochastic ISO Objective)}$$

Since it is impossible for the ISOto precisely model market players' demand without considering its strategic nature, optimization methods that are appropriate for deterministic and stochastic settings won't work here. Thus, as we specify in the next section, we model the market using RL.

C SMP

A typical structure of a market was based on the day-ahead market in which the ISOpredicts the following day's power demand and issues a *dispatch*, an offline production schedule to each producer while considering operational constraints and generation costs. The dispatch traditionally divides the 24-hour planning horizon into 48 discrete half-hour time periods. In addition to the generation of the predicted or nominal demand, the ISO also manages the reserve, which sets a backup production capability for each time step. If in real-time the controlled production determined by the dispatch is not enough to cover the realized demand, reserves, which are more flexible but also more expensive and polluting, are activated by an online controller. Producers are then compensated based on the System Marginal Price (SMP) mechanism, which is calculated as the marginal cost of producing the final unit of energy required to satisfy system demand, based on the least-cost dispatch solution.

Formally, let:

- P_t be the total power production at time t,
- D_t be the total system demand at time t,
- $C(P_t)$ be the cost function for production.

The SMP at timestamp t is defined as:

$$\kappa_t = \left. \frac{\partial C(P_t)}{\partial P_t} \right|_{P_t = D_t},$$

where κ_t represents the marginal cost of meeting the demand D_t using the least-cost generation defined by the merit-order curve.

In electricity markets, the SMP clears the market by equating supply and demand while satisfying the economic dispatch problem:

$$\min_{P_t} \; C(P_t) \quad \text{subject to} \quad P_t = D_t.$$

The SMP ensures that all dispatched generators receive the same price, incentivizing efficiency and cost-reflective bidding in competitive electricity markets. Note that SMP is non-decreasing with respect to the amount of power produced, meaning higher power demand usually results in a higher price *per kWh*. Consequently, reducing peak consumption is critical for lowering overall costs in the electricity market.

D Dynamic Programming Formulation for a Storage Only PCS-unit Agent

The dynamic programming formulation for the optimization problem for storage control is given as:

- **State Variables:**
  - Current time step t,
  - Current storage level σ_t.
- **Decision Variables:**
  - Energy bought P_t^b,
  - Energy sold P_t^s.
- **Transition Function:**

$$\sigma_{t+1} = \sigma_t + (P_t^b - l_t - P_t^s).$$

- **Objective Function:** The immediate reward at each time step is:

$$r(P_t^b, P_t^s) = \phi_t(P_t^s) - \xi_t(P_t^b).$$

  The cumulative reward is maximized over all time steps.
- **Recurrence Relation:**

$$V(t, \sigma_t) = \max_{P_t^b, P_t^s} \Big[r(P_t^b, P_t^s) + V(t+1, \sigma_{t+1}) \Big],$$

  subject to the constraints.

Similar methods adapted for stochastic optimization could be employed for the case where distribution is either known or can be approximated from existing data. In the case of the stochastic demand, there may even be an ability to compute a contingent policy that would deal with the stochastic signals.

E Quadratic Pricing

This example demonstrates the possible impact of price intervention on market dynamics. We assume deterministic setting for the ISOfor clarity, but the same logic can be applied in the non-deterministic scenario. Drawing from the literature Papadaskalopoulos & Strbac (2015), we apply superlinear and sublinear pricing adjustments to selling and feed-in tariffs, respectively.

The selling price incorporates a superlinear component:

$$\xi_t = \lambda^{buy} * P_t^b + \beta * [P_t^b]^2,$$

where λ^{buy} is a baseline price. Similarly, the feed-in price adds a sublinear adjustment:

$$\phi_t = \lambda^{feedin} * P_t^s + \gamma * \sqrt{P_t^s},$$

where λ^{feedin} is the baseline feed-in price.

Once per episode, at $t = 0$, the ISO commits to six coefficients $(\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2)$ that instantiate the quadratic tariff $\pi^{\text{buy/sell}}(x)$. These coefficients stay fixed for the ensuing T steps; dispatch tweaks δ_t may still follow online if enabled.

Baseline Scenario Assume the demand structure as described by Table 2 and $\rho = 0.3$. Also assume a single market player, operating a 30 kWh battery with charging/discharging limits of 30 kWh without internal load or generation capabilities. Under static prices ($\lambda^{buy} = \lambda^{feedin} =$ Baseline price, $\gamma = \beta = 0$), the optimal solution for the player is to charge fully at $t = 2$ and discharge fully at $t = 5$, yielding a profit of 4.5\$. Given this behavior, the ISO pays a cost of 138.75\$.

Timestamp	Baseline Price ($)	Base Demand (kWh)
1	0.40	40
2	0.35	35
3	0.40	40
4	0.45	45
5	0.50	60
6	0.45	45

Table 2: Baseline demand and prices

Impact of Price Intervention Now assume the ISO is willing to implement the intervention, and to set non-linear price signals. The ISO optimizes the price parameters, setting $\beta = 0.002$ and $\gamma = 0.455$ by solving for the objective function described in Equation ISO objective. This price adjustment incentivizes the player to redistribute charging and discharging activities, as the player solves the problem described in Section B.3. The optimal strategy for the player is as shown in Table 2, resulting in a higher profit of 6.52\$, including a subsidy from the ISOto the player (via sublinear feed-in price component) of 3.27\$. For the ISOtotal costs are reduced to 118.21\$ with the subsidy included. The intervention eliminates inefficiencies, benefiting both the ISOand the market player.

This example highlights the potential of price intervention to align market players' behavior with system-level efficiency goals. Furthermore, it demonstrates that the price intervention is not a zero-sum game, and some interventions can be beneficial for all parties involved.

However, What is described here is just one price intervention type possible. In general, the ISOwould explore the space of all possible price interventions to find the optimal one. We suggest searching in this space using RL methods.

F Day-Ahead Pricing as a Bandit Problem

At time 0, the ISO fixes prices in advance for all t, and receives a reward after the 48-timestep episode ends. This makes the ISO decide about the prices once per episode, which matches the dispatch decision. This turns the problem into a (very complex) bandit problem.

The bandit problem for dispatch and pricing in an electricity market is defined by the tuple:

$$\mathcal{B} = \langle \mathcal{A}, \mathcal{R}, \mathbb{P}, T \rangle$$

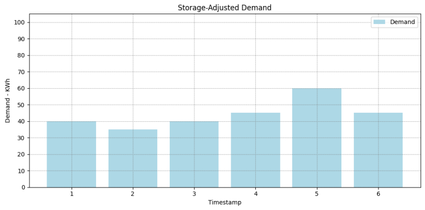

Figure 1: Original demand

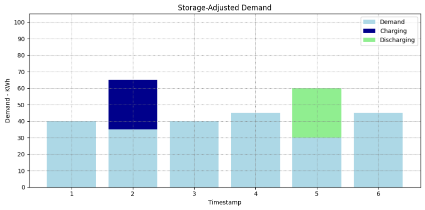

Figure 2: Linear Prices

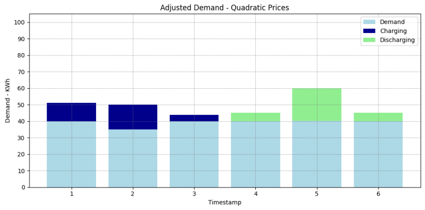

Figure 3: Quadratic Charging, Sublinear Discharging

Figure 3: Non-linear prices implementation

where:

- $\mathcal{A} = \{(d, p) \mid d \in \mathcal{D}, p \in \mathcal{P}\}$ is the set of actions, where each action is a pair (d, p):
  - $d \in \mathcal{D}$: Dispatch decision representing the amount of power to produce or allocate at a given time.
  - $p \in \mathcal{P}$: Price levels, including selling prices and feed-in tariffs offered to market participants.
- $\mathcal{R}(d, p)$ is the reward associated with selecting the action (d, p). Here, the reward is defined as the negative cost incurred by applying (d, p), such that:

$$\mathcal{R}(d, p) = -C(d, p),$$

  where $C(d, p)$ represents the total operational cost, including dispatch costs, market costs, and reserve activation costs.
- $\mathbb{P}(d, p)$ denotes the probability distribution governing the outcomes (e.g., market responses, demand realization) associated with the action (d, p).
- T is the time horizon, representing the total number of decision rounds.

At each time step $t \in \{1, 2, \dots, T\}$, the agent selects an action $(d_t, p_t) \in \mathcal{A}$, observes the resulting market dynamics and incurred cost $C(d_t, p_t)$, and receives a reward $\mathcal{R}(d_t, p_t) = -C(d_t, p_t)$.

The objective is to minimize the cumulative cost over the time horizon T, minimizing the cumulative regret R_T, defined as:

$$R_T = \sum_{t=1}^{T} C(d^*, p^*) - \sum_{t=1}^{T} \mathbb{E}[C(d_t, p_t)],$$

where (d^*, p^*) is the optimal dispatch and pricing policy that minimizes the expected cost:

$$(d^*, p^*) = \arg \min_{(d,p) \in \mathcal{A}} \mathbb{E}[C(d, p)].$$

This formulation addresses the trade-off between exploration (testing new dispatch and pricing strategies to learn their outcomes) and exploitation (applying strategies believed to minimize costs based on current knowledge).

G The Energy Market as MARL

In modeling modern power systems using marl, it is essential to account for multiple interacting perspectives. These include the physical constraints of the grid (e.g., stability limits), agent-level decision processes under partial observability, and the heterogeneity of demand profiles encompassing both nominal and flexible demand verify these are defined. Effective models must also incorporate market and pricing signals that influence agent behavior, and the temporal-spatial scalability required for real-world deployment. While these considerations are crucial for realistically and robustly capturing decentralized control strategies in complex energy environments, they also pose significant challenges to preserving the underlying Markovian structure that traditional agent-based decision models rely on.

G.1 Formal Model

Through the lens of RL, the ISO aims to learn an optimal policy that balances overall system efficiency with the mitigation of risk, such as insufficient power supply and grid instability. Simultaneously, market participants seek to maximize their individual utility in response to market signals, subject to their own operational constraints and preferences. We formally model this decentralized setting as a Markov Game (see Section 2), involving two types of agents: the ISO, and the GEAgents.

An important characteristic of the setting we aim to model is that the state space, action space, and reward functions are relatively straightforward to define. The complexity of solving this setting arises from modeling the joint transition function: the next state of the system and its stability depend on the actions performed by all agents.

Modeling the ISO

- **State Space $\mathcal{S}$:** Every time step t, typically representing a half-hour interval, the system state is associated with a vector $s_t \in \mathcal{S}$ that specifies operational factors that may affect decision-making. For the ISO this includes the system-level demand forecast $\hat{D}$ for the specified horizon, the system-level realized demand D_t for the current time step, supply capacities, storage states, etc. It may also include factors that indicate the stability state of the system, for example, whether the supply-demand balance is violated.
- **Action Space $\mathcal{A}$:** The ISO actions include the dispatch directives Δ_t that are given for each time step t and setting the sell prices $\xi_t(\cdot)$ and buy prices $\phi_t(\cdot)$ for each time step. In real-time the ISO also activates reserves and curtails power if needed, but assume these actions are dictated by the state and require no decision-making.

  Importantly, we support two types of pricing dynamics. In a day-ahead pricing regime, the ISO-makes the prices public at $t = 0$. In an online pricing setting, the ISO can dynamically set prices in response to the market signal. We discuss several pricing mechanisms and their characteristic, including the benefits of applying quadratic pricing, in Section 5.
- **Reward Function $\mathcal{R}$:** The ISO's reward integrates the economic efficiency and a risk measure to account for potential adverse outcomes arising from strategic GEAgents such that:

$$\mathcal{R} = -(C^{\text{dispatch}} + C_t^{\text{online}}) \qquad \text{(ISO objective)}$$

Modeling the GEAgents

- **State Space $\mathcal{S}$:** Each GEAgent is associated with a PCS-unit for which the state includes its local information (e.g., state-of-charge) as well as the price signal advertised by the ISO.
- **Action Space $\mathcal{A}$:** Modern GEAgents have significant decision-making autonomy, allowing them to choose how much energy to store, consume, or sell based on their local goals, capabilities, and constraints. In this work, we assume the GEAgent sees the current prices and its local state at the start of each iteration before deciding how to act. Also, both generation and consumption are non-controllable. Specifically, we only support generation via pv and consumption that is part of the non-flexible load of the PCS-unit. This means that generation and production are exogenous to the agent and are governed by a stochastic process, and the only decision variable is the charge and discharge actions, which may have stochastic effects.
- **Reward Function $\mathcal{R}$:** For each GEAgent i, the step-wise reward is the net cash flow obtained by trading with the grid:
$$\mathcal{R}_t^i \;=\; \phi_t\bigl(P_t^s, P_t^b\bigr) \;-\; \xi_t\bigl(P_t^s, P_t^b\bigr).$$
Maximising the cumulative sum of $\mathcal{R}_t^i$ over the horizon is equivalent to the strategic objective stated in (Strategic Player Objective), but written here without the expectation or the explicit time–index summation.

Joint Transition Function $\mathcal{T}$: Influence of Multiple Agents: Unlike a single-agent MDP, the Markov Game framework allows each agent's choice (including how GEAgents respond to prices or storage opportunities) to influence the next state. As mentioned above, the difficulty of modeling the transition function is at the core of the challenge. In general, the transition function can be decoupled into the state variables that are covered by the physical dynamics of the system. For example, when a charge or discharge action is performed, the battery dynamics obey:

$$\sigma_{t+1} \;=\; \sigma_t \;+\; \eta_{\mathrm{c}}\bigl[a_t\bigr]_+\,\Delta t \;-\; \eta_{\mathrm{d}}^{-1}\bigl[-a_t\bigr]_+\,\Delta t,$$

if an attempted action would violate $0 \le \mathrm{SoC} \le B_{\max}$ the short-fall or spillage is automatically settled with the grid, and a penalty is incurred. Propagated effect of local decisions, e.g., those solved with power flow.

Perhaps the most challenging aspect stems from the strategic interactions of the agents. These strategic decisions create a coupled system where each agent's payoff depends on the actions of others. In principle, the Markov Function $T(s_{t+1} \mid s_t, a_t^{ISO}, a_t^{PCS-unit})$ must fold together physical power flows, stochastic demand, renewables, battery chemistry and market clearing. Writing a closed-form T that captures all these layers is hopeless. Instead, we created the `Energy-Net` simulator (Section 4) maintain the physics and book-keeping, and we *learn* directly from roll-outs. This side-steps the need for explicit modeling of the complex dynamics and allows extracting value functions and policies using deep neural networks, rather than from first principles.

Episode As is typical in the day-ahead market, at the beginning of each episode (timestep $t = 0$) the ISO receives the predicated demand $\hat{D}$ for the next 48 half-hour intervals. It also receives the production and reserve capacities of its controlled units, the prices of each generated unit, and other information that might be relevant (i.e., weather forecast, special events, etc.). If day-ahead pricing is applied, the ISO sets and advertises the $\xi_t(\cdot)$ and feed-in tariff $\phi_t(\cdot)$ for the whole episode.

At each subsequent timestamp ($1 \le t \le 48$), the following sequence of events occurs:

1. The ISO observes the realized demand D_t.
2. If dynamic pricing is applied, the ISO sets the sell price $\xi_t(\cdot)$ and feed-in tariff $\phi_t(\cdot)$ for timestamp t.
3. The GEAgents strategically respond to the prices by buying or selling power to the grid.
4. If the net demand after accounting for the net power P_t^{net} exceeds the scheduled production (Δ_t), the ISO activates reserves (e.g., peaker plants) to cover the shortfall or curtails power to cover overloads.

This iterative process continues until the end of the planning horizon. Both agents seek a stationary (possibly stochastic) policy that maximizes their own long-term discounted accumulated reward.

H The `Energy-Net` Simulator

In spite of a variety of simulators that currently exist, there is no current framework that allows modeling the complex structure we want to account for and that is designed to work with off-the-shelf rl and marl methods. We therefore develop a novel simulator, `Energy-Net`[4], that we will use to examine our proposed solutions. `Energy-Net` is a modular, discrete–time simulator of a hybrid electricity market. The environment we develop is flexible and adaptable, and can be used to accommodate different system configurations. At the core of the design of the software is a decoupling between the physical dynamics of the electrical system and the strategic agents. `Energy-Net`is built around a strict *physics–agent split*. A high-fidelity physical core advances loads, renewables, batteries, and reserves, while the ISO and GEAgent interact only through a Gym-style `step()` interface. This design (i) lets us plug in any off-the-shelf rl/marl algorithm without touching the power-system code, (ii) isolates market rules in a single controller module, and (iii) ensures that learned policies can affect the grid *only* via explicit levers-prices and dispatch tweaks, thus preserving physical realism while streamlining experimentation.

Building on the formal setting introduced in Section 3, `Energy-Net` instantiates the 24-hour day-ahead electricity market. A single simulation episode therefore comprises T uniform intervals of length Δt (in our experiments T=48 and Δt=30min), together covering one 24-hour operational horizon. At each step $t \in \{1, \ldots, T\}$ the environment reveals the current forecast and grid state to the agents, applies their actions, propagates the physical dynamics, and returns next-state observations and rewards through the standard Gym `step` interface.

H.1 Physical Layer

Demand. System demand at each step is modelled as

$$D_t \;=\; f_{\text{seasonal}}(t) \;+\; \varepsilon_t,$$

where $f_{\text{seasonal}}(\cdot)$ captures the deterministic daily profile and $\varepsilon_t \sim \mathcal{N}\big(0, \sigma^2\big)$ is zero-mean Gaussian noise with user–configurable standard deviation σ.

GEAgent. Every PCS-unit hosts a single–block battery whose state of charge obeys

$$\sigma_{t+1} \;=\; \sigma_t \;+\; \eta_{\text{c}}\big[a_t\big]_+ \Delta t \;-\; \eta_{\text{d}}^{-1}\big[-a_t\big]_+ \Delta t,$$

subject to $0 \le \sigma_t \le S_{\max}$ and $|a_t| \le P_{\max}$. Here a_t is the charge (> 0) / discharge (< 0) power, $\eta_{\text{c}}, \eta_{\text{d}}$ are efficiency factors, and $P_{\max}$ the power limit.

Besides storage, each unit experiences *stochastic local load* l_t and PV generation g_t, drawn from configurable distributions. The net exchange with the grid is therefore

$$P_t^{\text{net}} \;=\; a_t \;+\; g_t \;-\; l_t.$$

Reserve. If $\Delta_t + P_t^{\text{net}} < D_t$, spinning reserve is activated and the simulator logs the penalty $C_t^{\text{reserve}}\big(D_t - \Delta_t - P_t^{\text{net}}\big)$, whose functional form and coefficients are user-configurable.

H.2 Market Layer

At each step t the ISO broadcasts a **buy tariff** $\phi_t(\cdot)$ (applied to energy flowing *into* storage) and a **sell tariff** $\xi_t(\cdot)$ (applied to energy flowing *out of* storage). `Energy-Net` supports two pricing regimes:

[4] link to repo - removed to respect the blind review process

a) **Online linear**. The operator chooses two bounded scalars $\lambda_t^{\text{buy}}, \lambda_t^{\text{sell}}$ and sets

$$\phi_t(P) = \lambda_t^{\text{buy}}, \qquad \xi_t(P) = \lambda_t^{\text{sell}}.$$

b) **Quadratic (*super-/sub-linear*)**. At the beginning of each episode ($t = 0$) the operator fixes four coefficients $\{\lambda^{\text{buy}}, \lambda^{\text{feedin}}, \beta, \gamma\}$; they remain unchanged for all subsequent steps. Power-dependent tariffs are then computed with exactly the same notation used in Section 5:

$$\xi_t = \lambda^{\text{buy}} P_t^b + \beta \left[P_t^b\right]^2, \tag{1}$$

$$\phi_t = \lambda^{\text{feedin}} P_t^s + \gamma \sqrt{P_t^s}. \tag{2}$$

Here β adds a *super-linear* surcharge to purchases, whereas γ grants a *sub-linear* bonus on injections. Optional real-time dispatch perturbations δ_t can still be issued on top of these pre-committed price curves.

H.2.1 Agent Interfaces

ISO observations. At each step t the operator receives $(t, \hat{D}_t, \widehat{P_t^{\text{net}}})$, where the hat denotes a one–step-ahead forecast of the aggregated exchange of all PCS-units.

PCS observations. Every storage unit observes the tuple $(t, \xi_t, \phi_t, \sigma_t)$.

ISO actions.

- *Online linear.* Set the instant tariff pair (ξ_t, ϕ_t) (+ optional dispatch tweak δ_t).
- *Quadratic (super-/sub-linear).* Commit the coefficient quadruple $(\lambda^{\text{buy}}, \lambda^{\text{feedin}}, \beta, \gamma)$ that parameterises; these remain fixed for the whole episode.

PCS action. A single continuous decision $a_t \in [-D_{\max}, C_{\max}]$ interpreted as charge $[a_t > 0]$ or discharge $[a_t < 0]$.

H.2.2 Reward Structure

Per-step rewards follow the definitions already introduced in Section B.4.

H.2.3 Multi–Agent Execution

Energy–Net wraps both agents in a *single* multi–agent environment that extends the GYMNASIUM interface Towers et al. (2024). `step(...)` consumes a dictionary of actions and returns observation, reward, and termination tuples keyed by agent identity. Internally, a unified `EnergyNetController` advances the simulation in the following sequential order:

1. **Price setting —** the ISOchooses tariffs (and, if enabled, dispatch).
2. **Battery control —** the PCS-unitresponds with its charge or discharge command.
3. **Energy exchange —** supply, demand, and storage flows are balanced; any shortfall triggers spinning reserve.
4. **State update and reward —** physical states, SoC, and financial ledgers are updated, and rewards are computed for both agents.

This integrated design eliminates manual data transfer between separate environments and exposes consistent, step–level metrics for training and evaluation. Notably, additional assets — renewables, alternative storage chemistries, custom reward definitions — can be introduced by registering new modules that comply with the interfaces above; no modification of the core simulation loop is required.

I Evaluation Setup

Table 3: Scenario matrix used throughout Section 6. Columns 2–3 describe the **ISO** policy elements; column 4 the **PCS**. “Learned” means the dispatch network is frozen from the previous scenario while the remaining degrees of freedom are (re-)trained with TD3.

ID	ISO pricing	ISO dispatch	PCS behaviour
`Baseline`	N/A	Equal to predicted demand	N/A
`ISO-Dispatch`	N/A	Learned	N/A
`ISO-L`	Online linear	Learned (prior S2)	Deterministic / fixed
`ISO-Q`	Quadratic	Learned (prior S2)	Deterministic / fixed
`Joint-Storage-L`	Online linear	Learned (prior S3)	**Learned**
`Joint-Storage-Q`	Quadratic	Learned (prior S3)	**Learned**
`Joint-PCS-L`	Online linear	Learned (prior S4)	**Learned + intrinsic load/production**
`Joint-PCS-Q`	Quadratic	Learned (prior S4)	**Learned + intrinsic load/production**

Global scenario parameters (all baselines).

- *Demand pattern: sinusoidal*

$$D_t \;=\; L_0 \;+\; A\cos\!\Big(\tfrac{2\pi}{P}\,(kt+\phi)\Big)$$

  with base load $L_0 = \mathbf{150}$ MWh, amplitude $A = \mathbf{50}$ MWh, interval multiplier $k = \mathbf{8}$, phase shift $\phi = \mathbf{5}$, period divisor $P = \mathbf{24}$.
- Dispatch energy price: \$100 per MWh.
- Reserve energy price: \$300 per MWh.
- Forecast-error noise (prediction error): $\sigma = \mathbf{10}$ MWh.

For each interval t we first sample the *realised* demand D_t from the sinusoidal profile above. The ISO observes only a noisy one-step-ahead prediction

$$\hat{D}_t \;=\; D_t \;+\; \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}\big(0, \sigma^2\big).$$

Hence, each experiment measures both the forecast error and the operator’s reaction to it. Note that even in the *day-ahead* pricing scenarios, where the six tariff coefficients chosen at $t = 0$ remain fixed throughout the episode, the instantaneous ISO reward r_t^{ISO} is still computed *at every step*. This preserves time-resolved feedback while respecting the regulatory commitment to day-ahead prices.

J Results

Local context re-activates storage.
Without an intrinsic load/production signal (`Joint-Storage-L` & `Joint-Storage-Q`) the battery never exchange energy and the column *PCS-unit Exchange* in Table 4 is 0 MWh. Introducing even a modest prosumer profile (`Joint-PCS-L` & `Joint-PCS-Q`) forces the unit to interact with the grid, shifting about 442 MWh (linear tariff) or 324 MWh (quadratic tariff) over the 48-step episode.

Reserve energy is largely supplanted.
The extra flexibility supplied by the battery allows the ISOto rely less on spinning reserve: the quantity drawn falls from 176 MWh (`ISO-L`) and 121 MWh (`ISO-Q`) down to 117 MWh in the

Table 4: Episode-total *cost* and *energy* breakdown across all evaluated scenarios (see Table 3 for scenario definitions).

Scenario	**Dispatch**		**Reserve**		**PCS-unit Exchange**
	Cost [$]	Energy [MWh]	Cost [$]	Energy [MWh]	Energy [MWh]
`Baseline`	720 000 ± 10	7 200 ± 0.1	62 400 ± 30	208 ± 0.1	0
`ISO-L`	728 235.04 ± 5 089.24	7 282.34 ± 50.89	52 815 ± 5 721	176.05 ± 19.07	800 ± 0
`ISO-Q`	750 698.08 ± 3 502.32	7 506.98 ± 35.02	36 321 ± 1 134	121.07 ± 3.78	800 ± 0
`Joint-Storage-L`	812 603.1 ± 1 071.45	8 126.13 ± 1.07	44 400 ± 282	148 ± 0.94	0
`Joint-Storage-Q`	812 621.48 ± 1 012.64	8 126.21 ± 1.01	44 400 ± 318	148 ± 1.06	0
`Joint-PCS-L`	732 244.02 ± 3 602.57	7 322.44 ± 36.02	50 541 ± 1 242	168.47 ± 4.14	442.14 ± 9.61
`Joint-PCS-Q`	745 062.53 ± 343.37	7 450.62 ± 36.43	35 100 ± 612	117 ± 2.04	324 ± 8.40

quadratic joint scenario. Because reserve blocks are the most carbon and price intensive resource, substituting them with stored energy directly improves both sustainability and operating margins.

Quadratic pricing yields the best balance.
Relative to the online linear tariff, the quadratic day-ahead curve cuts reserve usage by ≈30 % with only 324 MWh of battery throughput (cf. 442 MWh under the linear scheme). The slight 128 MWh increase in scheduled dispatch is more than offset by the smaller reserve call and lower battery wear.

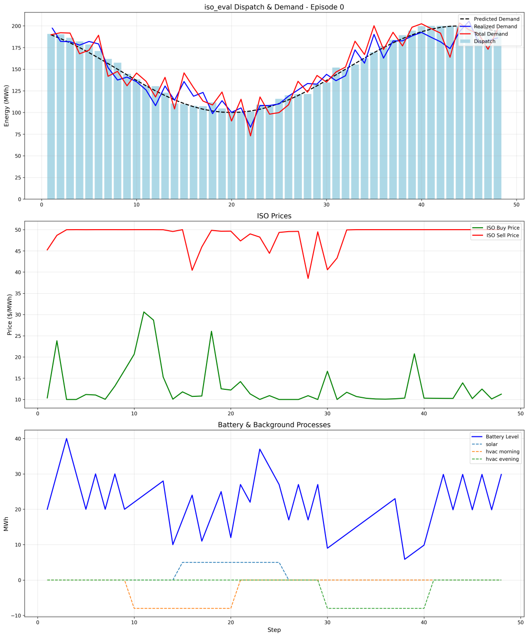

Figure 4: Energy, tariff, and cost traces for **`Joint-PCS-L`**. Top: dispatch vs. realised demand. Middle: ISO buy/sell tariff trajectories. Bottom: cumulative cost distribution at episode end.

J.1 Empirical Evaluation Process

J.1.1 `Baseline` – Fixed Day-Ahead Schedule

This baseline freezes the ISO's day-ahead schedule at the one-step demand forecast and publishes *no* real-time prices, so the PCS-unit stays idle. Across the 48 × 30 min horizon the grid delivers **7 200 MWh** of scheduled generation and calls **208 MWh** of spinning reserve, with *zero* battery exchange. These figures serve as the reference for all percentage comparisons that follow.

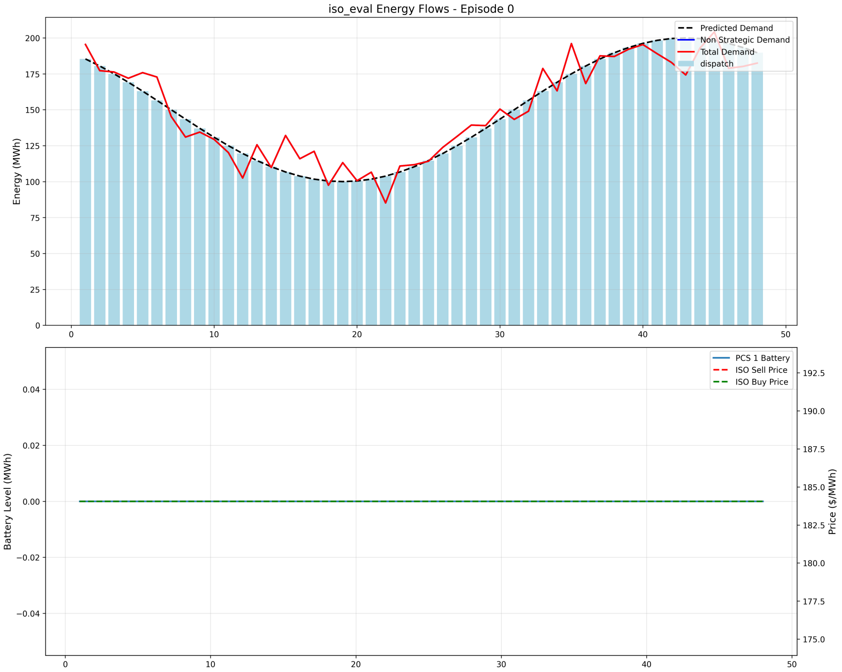

Figure 5: Energy-flow profile for **Baseline**. *Top:* dashed = forecast demand, solid red = realised demand, blue bars = fixed day-ahead dispatch. *Bottom:* battery state of charge stays at 0 and buy/sell prices coincide, confirming the absence of storage actions or dynamic tariffs.

J.1.2 ISO-Dispatch – Adaptive Dispatch, No Price Signal

In this scenario the ISOcan *revise the dispatch level* every 30 minutes to track its demand forecast, but it still publishes no real-time prices, so the PCS-unit remains idle. The configuration isolates the pure value of feed-forward unit-commitment.

Relative to the fixed day-ahead baseline (Baseline):

- Scheduled generation rises from 7 200 MWh to **7 229 MWh** +0.4 %.
- Reserve energy increases from 208 MWh to **249 MWh** +19 %.

The extra 29 MWh of dispatch more than offsets the reserve reduction, showing that unit-commitment alone cannot handle real-time variability efficiently when no flexible resource is available.

J.1.3 ISO-L – Linear Price Signal with *Pre-defined* PCS Actions

In this variant the ISO updates its dispatch each half-hour and also posts a real-time *linear* buy/sell tariff. The PCS-unit , however, does **not** react; it follows an offline schedule that charges during the early-morning valley and discharges at the evening peak. All storage moves are therefore deterministic and price-agnostic.

Key energy effects relative to the fixed baseline (Baseline):

- Battery activity - The preset cycle moves *800 MWh* from low-demand to high-demand hours (Table 4, last column).
- Scheduled generation - Dispatch rises from 7 200 MWh to *7 282 MWh* (+1.1 %).
- Reserve usage - Spinning reserve falls from 208 MWh to *176 MWh* (–15 %).

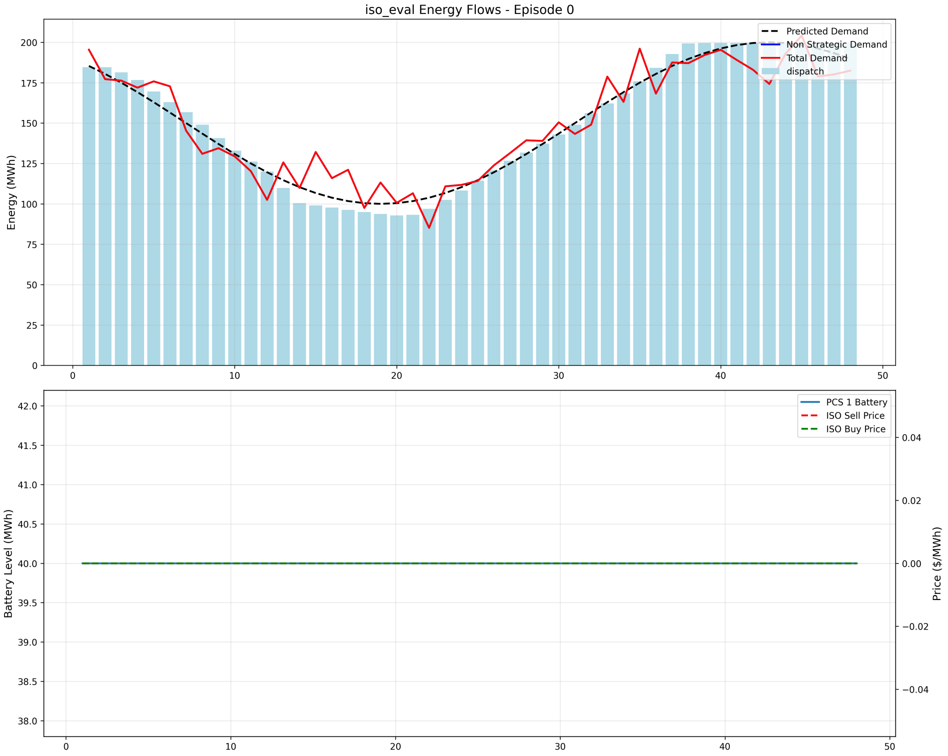

Figure 6: Energy–flow profile for **ISO-Dispatch**. *Top:* dashed = forecast demand, solid red = realised demand, blue bars = adaptive dispatch. *Bottom:* battery state of charge remains at 0 and the buy/sell tariff is flat, confirming that no storage actions or dynamic prices are present.

The fixed cycle smooths the net load enough to cut reserve energy by 32 MWh, but that benefit is partly offset by an 82 MWh increase in scheduled generation. In short, a pre-programmed battery can firm the load profile, yet it is still less effective than a storage agent that responds optimally to real-time prices.

J.1.4 `ISO-Q` – Quadratic Price Signal with Pre-defined PCS Actions

The ISO now publishes a *quadratic* buy/sell tariff (three coefficients per side) while the PCS-unit still follows the fixed charge–discharge cycle of `ISO-L`.

Energy impact relative to the fixed baseline (`Baseline`):

- Battery activity - unchanged at *800 MWh* (preset cycle).
- Scheduled generation - rises to *7 507 MWh*, an increase of 307 MWh (+4.3
- Reserve usage - falls to *121 MWh*, a 42% drop versus 208 MWh in `Baseline`and a further 31 % reduction compared with the linear-price case (`ISO-L`).

Quadratic pricing therefore achieves the *lowest* reserve energy of all pre-defined scenarios, even though the battery does not react to prices, by letting the ISO shape its real-time tariff more aggressively around the deterministic storage profile.

J.2 `Joint-Storage-L` and `Joint-Storage-Q` – TD3 ISO, Learned PCS

In these scenarios both agents are trained with TD3. The controls real-time prices and dispatch; the PCS-unit is now free to learn its own policy. Regardless of whether the tariff is linear or quadratic,

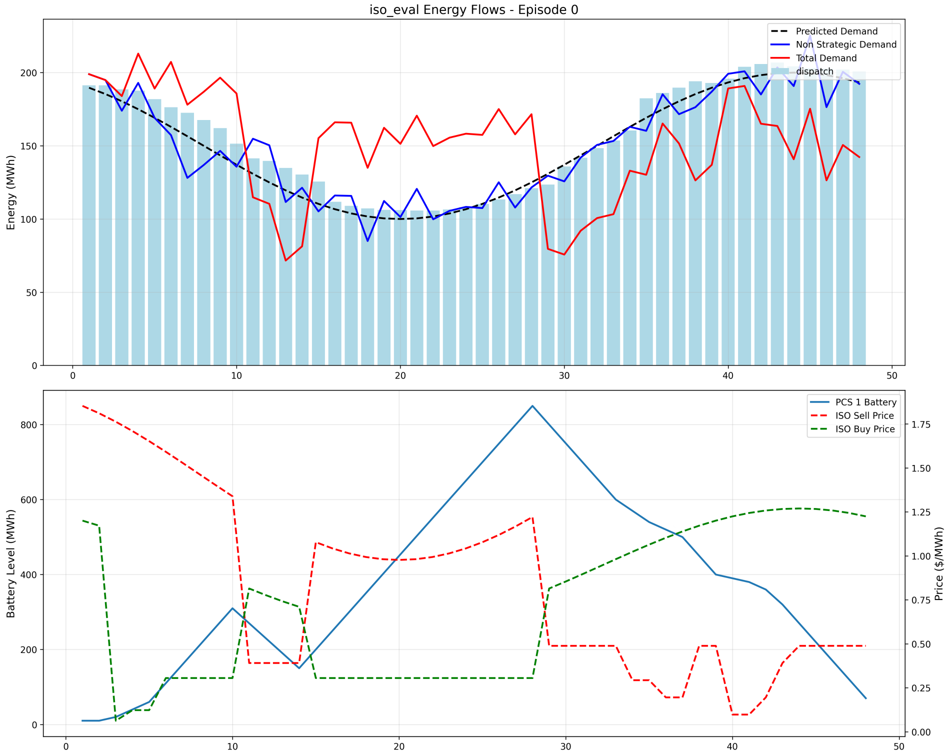

Figure 7: Energy–flow profile for **ISO-Q** (quadratic prices, deterministic PCS).

the quickly discovers that posting the *maximum* allowed buy/sell price removes any profitable arbitrage. The learned PCS-unit therefore chooses to stay idle, and the battery never moves energy.

Key energy outcome (identical for L and Q):

- Battery exchange - *0 MWh*.
- Scheduled generation - *8 126 MWh* (+13% versus the 7 200 MWh baseline).
- Reserve usage - *148 MWh* (–29% relative to 208 MWh in `Baseline`).

Key observation. By exploiting its price-setting power the captures all potential surplus, pushing the system into a “monopolistic” equilibrium that eliminates storage activity. Reserve demand does fall, but only at the cost of a large increase in base-load dispatch; the grid loses the flexibility benefit that an active battery would provide.

J.3 `Joint-PCS-L` – Learned ISO and PCS under Endogenous Load & Production

The fully learned setting of Section `Joint-Storage-L` and `Joint-Storage-Q` collapsed into a “monopolistic” equilibrium because the storage unit had *no reason* to transact. To restore economic pressure we embed the PCS-unit in a simple prosumer model:

- **Background HVAC load** – square-wave, 8 kW peak.
- **Rooftop PV** – bell-shaped profile, 5 kW peak at solar noon.

Whenever the net local balance is negative the battery must buy from the grid; when positive it can inject. Both and PCS-unit continue to train with TD3, and the ’s tariff are *online linear* or *quadratic*.

Energy outcomes (from Table 4):

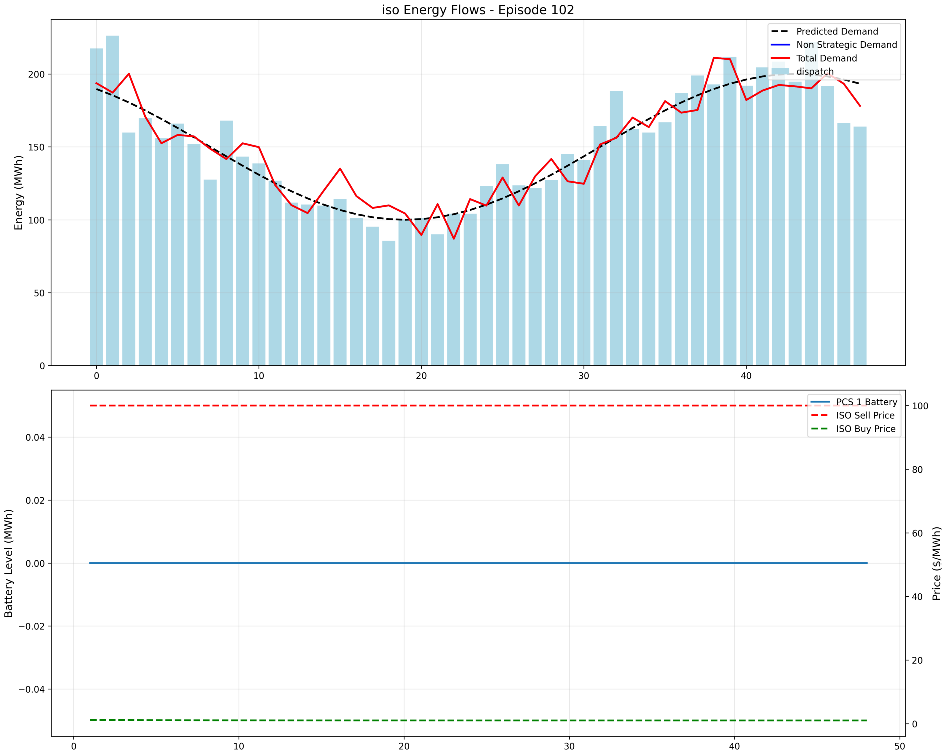

Figure 8: Energy–flow profile for **`Joint-Storage-L`**. The ISO posts buy/sell tariffs at their upper limit, leaving the battery inactive (0 MWh exchange).

- Battery exchange - *442 MWh* shuffled across the day (30% of steps involve a charge or discharge).
- Scheduled generation – *7 322 MWh* (very close to the deterministic baseline).
- Reserve usage – *168 MWh*, midway between the linear pre-defined case (176 MWh) and the best quadratic case (117 MWh).

Endogenous prosumer dynamics "wake up" the battery: facing real cost when HVAC load peaks and real revenue when PV over-produces, the agent learns to arbitrage once again. The adapts by moderating its price ceiling: tariffs remain high enough to steer the battery but no longer saturate at the upper bound, breaking the deadlock observed in `Joint-Storage-L`.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper. Only add this information once your submission is accepted and deanonymized. The acknowledgments do not count towards the 8–12 page limit.

References

Victor Ahlqvist, Pär Holmberg, and Thomas Tangerås. A survey comparing centralized and decentralized electricity markets. *Energy Strategy Reviews*, 40:100812, 2022.

Stefano V. Albrecht, Filippos Christianos, and Lukas Schäfer. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2024. URL https://www.marl-book.com.

Flora Charbonnier, Thomas Morstyn, and Malcolm D McCulloch. Coordination of resources at the edge of the electricity grid: Systematic review and taxonomy. *Applied Energy*, 318:119188, 2022.

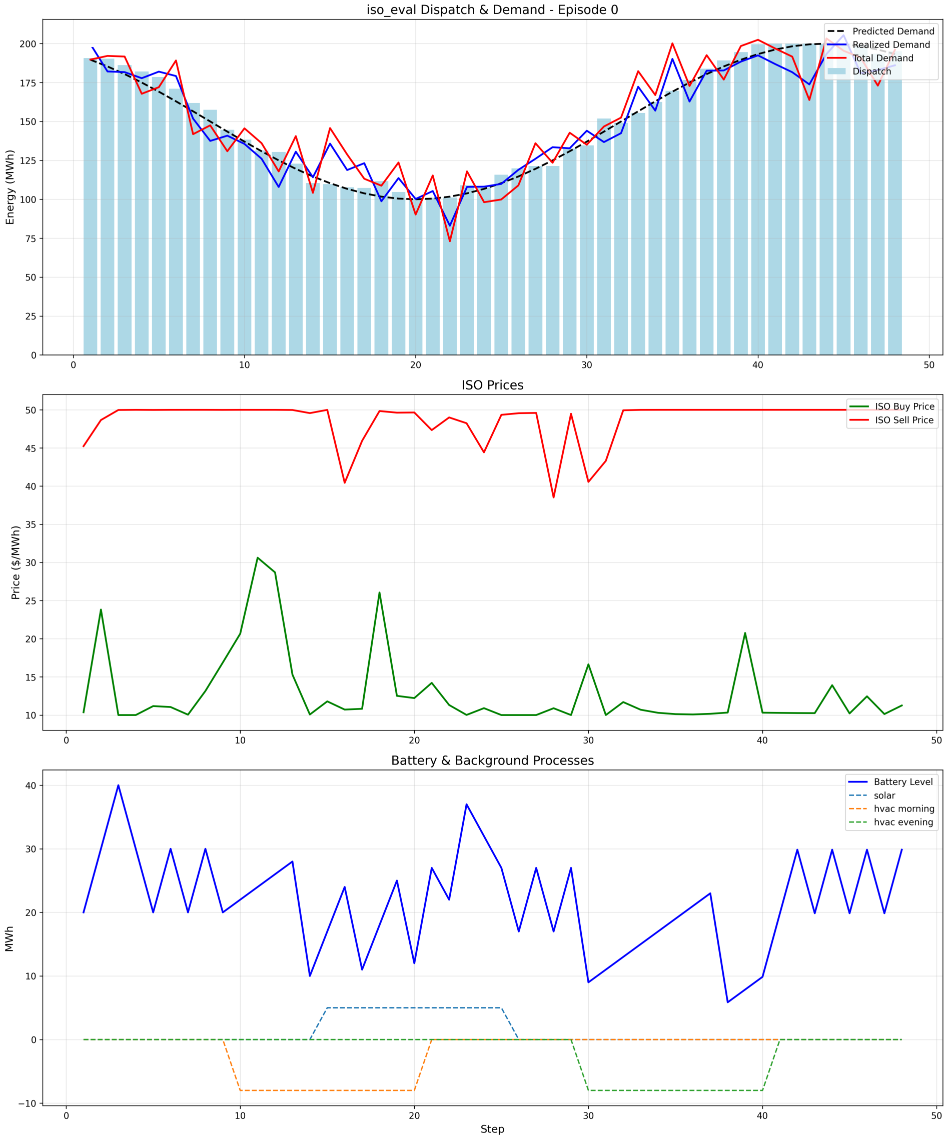

Figure 9: Energy, tariff and battery SoC traces for **Joint-PCS-L**. The learned PCS cycles 442 MWh in response to its own load/PV profile and ISO prices, cutting reserve demand to 168 MWh.

Elinor Ginzburg-Ganz, Itay Segev, Alexander Balabanov, Elior Segev, Sivan Kaully Naveh, Ram Machlev, Juri Belikov, Liran Katzir, Sarah Keren, and Yoash Levron. Reinforcement learning model-based and model-free paradigms for optimal control problems in power systems: Comprehensive review and future directions. *Energies*, 17(21):5307, 2024. ISSN 1996-1073. DOI: 10.3390/en17215307. URL https://doi.org/10.3390/en17215307.

Chenxiao Guan, Yanzhi Wang, Xue Lin, Shahin Nazarian, and Massoud Pedram. Reinforcement learning-based control of residential energy storage systems for electric bill minimization. In *2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 637–642. IEEE, 2015.

Nick Harder, Anke Weidlich, and Philipp Staudt. Finding individual strategies for storage units in electricity market models using deep reinforcement learning. *Energy Informatics*, 6(Suppl 1):41, 2023. DOI: 10.1186/s42162-023-00293-0.

Antoine et al. Marot. Learning to run a power network challenge: a retrospective analysis. In *NeurIPS 2020 Competition and Demonstration Track*, 2021.

Panagiotis Michailidis, Iakovos Michailidis, and Elias Kosmatopoulos. Reinforcement Learning for Optimizing Renewable Energy Utilization in Buildings: A Review on Applications and Innovations. *Energies*, 18(7):1724, 2025. DOI: 10.3390/en18071724.

Takao et al. Moriyama. Reinforcement learning testbed for power-consumption optimization. In *Methods and Applications for Modeling and Simulation of Complex Systems: 18th Asia Simulation Conference (AsiaSim)*. Springer, 2018.

Aviad Navon, Juri Belikov, Ariel Orda, and Yoash Levron. On the stability of strategic energy storage operation in wholesale electricity markets. *arXiv preprint arXiv:2402.02428*, 2024. URL https://arxiv.org/abs/2402.02428.

Dimitrios Papadaskalopoulos and Goran Strbac. Nonlinear and randomized pricing for distributed management of flexible loads. *IEEE Transactions on Smart Grid*, 7(2):1137–1146, 2015.

ATD Perera and Parameswaran Kamalaruban. Applications of reinforcement learning in energy systems. *Renewable and Sustainable Energy Reviews*, 137:110618, 2021.

Aisling Pigott, Constance Crozier, Kyri Baker, and Zoltan Nagy. Gridlearn: Multiagent reinforcement learning for grid-aware building energy management. *Electric Power Systems Research*, 2022.

Xin Qiu, Tu A Nguyen, and Mariesa L Crow. Heterogeneous energy storage optimization for microgrids. *IEEE Transactions on Smart Grid*, 7(3):1453–1461, 2015.

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 1953.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 1:1–8, 2024.

José R Vázquez-Canteli and Zoltán Nagy. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied energy*, 235:1072–1089, 2019.

José R. Vázquez-Canteli, Jérôme Kämpf, Gregor Henze, and Zoltan Nagy. Citylearn v1.0: An openai gym environmfent for demand response with deep reinforcement learning. Association for Computing Machinery, 2019.

Lucien Werner and Peeyush Kumar. Multi-market energy optimization with renewables via reinforcement learning. *arXiv preprint arXiv:2306.08147*, 2023. URL https://arxiv.org/abs/2306.08147.

Thomas Wolgast and Astrid Nieße. Approximating energy market clearing and bidding with model-based reinforcement learning. *arXiv preprint arXiv:2303.01772*, 2023. URL https://arxiv.org/abs/2303.01772.

Ting Yang, Liyuan Zhao, Wei Li, and Albert Y Zomaya. Dynamic energy dispatch strategy for integrated energy system based on improved deep reinforcement learning. *Energy*, 235:121377, 2021.

Bin Zhang, Weihao Hu, Di Cao, Qi Huang, Zhe Chen, and Frede Blaabjerg. Deep reinforcement learning–based approach for optimizing energy conversion in integrated electrical and heating system with renewable energy. *Energy conversion and management*, 202:112199, 2019.

Ziqing Zhu, Ze Hu, Ka Wing Chan, Siqi Bu, Bin Zhou, and Shiwei Xia. Reinforcement learning in deregulated energy market: A comprehensive review. *Applied Energy*, 345:120360, 2023.

Supplementary Materials

The following content was not necessarily subject to peer review.

Content that appears after the references are not part of the "main text," have no page limits, are not necessarily reviewed, and should not contain any claims or material central to the paper. If your paper includes supplementary materials, use the

```
\beginSupplementaryMaterials
```

command as in this example, which produces the title and disclaimer above. If your paper does not include supplementary materials, this command can be removed or commented out.