Multi-Agent Reinforcement Learning for Optimal Bidding of Energy Aggregators in Real-Time Multi-Service Provision

Truong Hoang Bao Huy, Hang Nguyen, Kha T. Nguyen, Christina Papadimitriou, Phuong H. Nguyen
Department of Electrical Engineering
Eindhoven University of Technology
Eindhoven, The Netherlands

***Abstract*—The growing penetration of renewable energy generation and electrification, along with the rise of energy aggregators, necessitates advanced decision-support frameworks that can coordinate distributed flexibility across multiple electricity services. This paper proposes a two-stage multi-service scheduling framework for an aggregator that participates in day-ahead energy and reserve markets while providing real-time balancing service. In Stage 1, a model-based optimization co-optimizes day-ahead energy and reserve bids. In Stage 2, real-time dispatching is modeled as a partially observable Markov game and solved using an attention-based multi-agent deep deterministic policy gradient (ATT-MADDPG) algorithm. Case studies on an aggregator of four prosumers, validated on real Dutch prosumer and market data, show that the ATT-MADDPG effectively adapts to uncertainty by coordinating prosumer flexibility across reserve and balancing markets.**

***Index Terms*—Energy aggregator, multi-agent reinforcement learning, multi-service provision, balancing service, real-time dispatching**

I. Introduction

Recent advances in smart home technologies, including home energy management systems (HEMSs), sensors, and smart appliances, promise to transform formerly passive and inflexible consumers into active and flexible prosumers. In this emerging landscape, the aggregator acts as an energy supplier and service provider that represents its customers by pooling flexibility from numerous small assets into a sufficiently large and controllable portfolio, thereby enabling participation in various markets, including the day-ahead, reserve, and balancing markets. However, coordinating the heterogeneous portfolio under significant uncertainty remains highly challenging, particularly for reserve dispatch and balancing services that require real-time responsiveness [1]. Therefore, this paper proposes an intelligent real-time dispatching strategy that enables optimal utilization of flexible assets, maximizes aggregator profit, and enhances overall system balancing.

Model-based optimization approaches have been widely employed to investigate aggregator bidding strategies for providing services across multiple markets, including reserve capacity [2]–[5], redispatch [6], balancing service [7], and voltage/frequency regulation [8], [9]. Given the high uncertainties in prosumer demand, renewable generation, and market prices, *model-based* optimization methods typically address risk using stochastic programming [2], [3], robust optimization [5], [7], or distributionally-robust optimization [8]. However, these methods either rely on a limited set of representative scenarios or give overly conservative decisions. To handle real-time uncertainties, model predictive control (MPC) [4] offers some robustness, yet it is still largely constrained to single-agent or centralized models and incurs a high computational burden. In general, *model-based* approaches presume access to precise probability distributions or highly accurate forecasts of uncertain parameters, which is difficult to achieve in inherently stochastic and dynamic real-world environments [1].

Reinforcement learning (RL) is a *model-free* approach for sequential, dynamic control in which agents learn optimal decisions through continuous interaction with the environment, without prior knowledge of the system. RL algorithms are well suited to real-time energy scheduling as their policies execute in milliseconds without further system identification. Prior works have applied various single-agent RL (SARL) methods to solve the optimal bidding strategy for aggregators, including deep Q-learning (DQN) for EV-based frequency regulation (FR) [10], deep deterministic policy gradient (DDPG) for voltage regulation [11], [12]. However, SARL suffers from the curse of dimensionality and non-stationarity induced by other agents when scaling to problems involving a large number of agents. To mitigate these issues, recent studies have turned to multi-agent RL (MARL), such as voltage regulation using multi-agent deep deterministic policy gradient (MADDPG) [13] and multi-agent actor-critic (MAAC) [14], and frequency reserve and voltage regulation using multi-agent proximal policy optimization (MAPPO) [15]. The use of MARL for multi-service provision problems remains limited. Moreover, existing MARL methods either rely on centralized joint policies or adopt centralized-training–decentralized-execution (CTDE) critics that simply concatenate observations and actions of all agents, resulting in a critic dimension that grows rapidly and making training difficult to scale to many agents.

Motivated by the above literature review, this study proposes a two-stage optimal bidding strategy that enables an aggregator of prosumers to jointly deliver multiple market products by coordinating flexibility assets. The main contributions are summarized as follows:

- Formulate the real-time ancillary service provision as a Partially Observable Markov Game (POMG) that captures decentralized decision-making, local observations, and interactions among prosumers, and design a fully model-free MARL solution that learns operational policies directly from interaction data.

This research was funded by the CETPartnership, under the Joint Call 2023. The CETPartnership's research projects are co-funded by the European Commission (Grant Agreement No. 101069750) and the national funding organizations listed on the CETPartnership website, Project No. CETP-2023-00493 (DigiRES) https://www.digi-res.eu/.

979-8-3315-8138-1

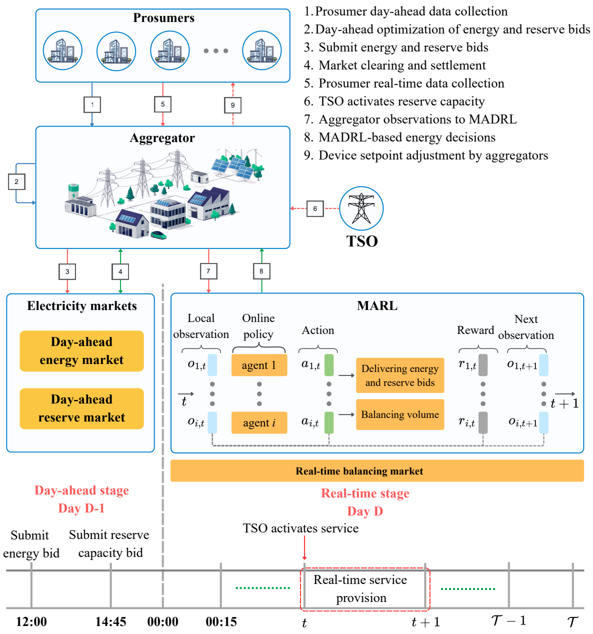

Fig. 1. Schematic view of the proposed two-stage framework.

- Develop an Attention-based Multi-Agent Deep Deterministic Policy Gradient (ATT-MADDPG) algorithm with decentralized continuous-control policies and an attention-enhanced centralized critic that focuses on the most relevant agents, providing a scalable approximation of the joint policy and improving training performance.
- Validate the proposed strategy on real Dutch prosumer data to demonstrate its effectiveness.

II. Problem Formulation

This study focuses on the multi-service provision by an aggregator that coordinates a prosumer portfolio willing to use their battery energy storage system (BESS) and PV system to participate in day-ahead energy and reserve markets while providing real-time balancing. As shown in Fig. 1, a two-stage framework is proposed: a model-based day-ahead optimization that co-optimizes energy and reserve bids, and a MARL-based real-time control layer that coordinates prosumer flexibility to deliver the contracted services.

A. Day-ahead scheduling

The objective is to maximize the profit of the aggregator from energy bidding and reserve capacity provision while accounting for BESS degradation costs:

$$\max \sum_{t\in\mathcal{T}} \Bigg[\Big(E_t^{da,+}(\lambda_t^{da} - \lambda^{p,+}) + E_t^{da,-}(\lambda^{p,-} - \lambda_t^{da}) \Big) + \lambda_t^R (\tilde{R}_t^{\uparrow} + \tilde{R}_t^{\downarrow}) - \lambda_t^{deg} \Delta t \sum_{i\in\mathcal{I}} \Big(P_{i,t}^{bess,c} \eta_i^{bess,c} + \frac{P_{i,t}^{bess,d}}{\eta_i^{bess,d}} \Big) \Bigg] \tag{1}$$

In (1), the first term is the energy margin of the aggregator between wholesale price λ_t^{da} and retail prices $\lambda^{p,\pm}$ for day-ahead imports $E_t^{da,-}$ and exports $E_t^{da,+}$ to supply demand or export surplus of prosumers. The second term is the revenue from upward and downward reserve capacities $\tilde{R}_t^{\uparrow/\downarrow}$ at price λ_t^R, and the last term penalizes BESS degradation.

The following constraints need to be satisfied:

$$P_{i,t}^{pv} + \tilde{R}_{i,t}^{pv,\uparrow} \le \hat{P}_{i,t}^{pv}, \quad P_{i,t}^{pv} - \tilde{R}_{i,t}^{pv,\downarrow} \ge 0, \quad \forall i,t \tag{2}$$

$$E_{i,t+1}^{bess} = E_{i,t}^{bess} + \Delta t \left(\eta_i^{bess,c} P_{i,t}^{bess,c} - \frac{P_{i,t}^{bess,d}}{\eta_i^{bess,d}} \right), \quad \forall i,t \tag{3}$$

$$\underline{E}_i^{bess} \le E_{i,t}^{bess} \le \overline{E}_i^{bess}, \quad \forall i,t \tag{4}$$

$$0 \le P_{i,t}^{bess,c} \le u_{i,t}^{bess} \overline{P}_i^{bess,c}, \quad \forall i,t \tag{5}$$

$$0 \le P_{i,t}^{bess,k} \le (1 - u_{i,t}^{bess}) \overline{P}_i^{bess,k}, \quad \forall i,t \tag{6}$$

$$\tilde{R}_{i,t}^{bess,\uparrow/\downarrow} \le \overline{P}_i^{bess,d/c} - P_{i,t}^{bess,d/c}, \quad \forall i,t \tag{7}$$

$$E_t^{da} = E_t^{da,-} - E_t^{da,+} = \sum_{i\in\mathcal{I}} E_{i,t}^{p,da}, \quad \forall t \tag{8}$$

$$E_{i,t}^{p,da} = \Delta t \left(P_{i,t}^{pv} + P_{i,t}^{bess,d} - P_{i,t}^{bess,c} - \hat{P}_{i,t}^{l} \right), \quad \forall i,t \tag{9}$$

$$\tilde{R}_t^{\uparrow/\downarrow} = \sum_{i\in\mathcal{I}} \tilde{R}_{i,t}^{p,\uparrow/\downarrow} \le \sum_{i\in\mathcal{I}} \left(\tilde{R}_{i,t}^{bess,\uparrow/\downarrow} + \tilde{R}_{i,t}^{pv,\uparrow/\downarrow} \right), \quad \forall t \tag{10}$$

Constraint (2) limits the upward and downward PV reserve capacities $\tilde{R}_{i,t}^{pv,\uparrow}$ and $\tilde{R}_{i,t}^{pv,\downarrow}$ based on the forecast available power $\hat{P}_{i,t}^{pv}$ and scheduled PV power $P_{i,t}^{pv}$. (3) updates the state of charge (SOC) of BESS, (4)–(6) enforces SOC and charge/discharge limits and avoids simultaneous charging and discharging using binary variable $u_{i,t}^{bess}$, and (7) defines the remaining headroom for reserve provision. (8) and (9) present the day-ahead energy of the aggregator E_t^{da} and each prosumer $E_{i,t}^{p,da}$ at the grid connection point, where $\hat{P}_{i,t}^{l}$ is the forecast inflexible load. (10) aggregates the upward and downward reserve capacities of prosumers into the day-ahead reserve bids of the aggregator $\tilde{R}_t^{\uparrow/\downarrow}$.

The outputs of day-ahead stage are the prosumer-level energy $E_{i,t}^{p,da}$ and reserve capacities $\tilde{R}_{i,t}^{p} = \tilde{R}_{i,t}^{p,\uparrow} - \tilde{R}_{i,t}^{p,\downarrow}$, together with the aggregator-level energy E_t^{da} and reserve commitments $\tilde{R}_t = \tilde{R}_t^{\uparrow} - \tilde{R}_t^{\downarrow}$, which serve as inputs to the real-time stage.

B. Real-time dispatching

In the second stage, the aggregator performs real-time dispatching to deliver the reserve capacity contracted in the day-ahead stage, minimize imbalance costs caused by forecast errors, while providing additional balancing services. This problem is modeled as a decentralized POMG:

1) Agents: Each prosumer $i \in \mathcal{I}$ is modeled as an individual agent that controls its own real-time decisions.

2) Observation: The environment state $s_t \in \mathcal{S}$ is represented by the joint set of agent observations $\{o_{1,t}, \dots, o_{\mathcal{I},t}\}$. For each agent i, the local observation $o_{i,t}$ is defined as:

$$o_{i,t} = \left[E_{i,t}^{p,da}, \tilde{R}_{i,t}^{p}, \overline{P}_{i,t}^{l,rt}, \overline{P}_{i,t}^{pv,rt}, E_{i,t}^{bess}, \lambda_t^{\uparrow}, \lambda_t^{\downarrow}, reg_t \right], \tag{11}$$

(11) includes: (1) day-ahead energy bid $E_{i,t}^{p,da}$ and reserve capacity bid $\tilde{R}_{i,t}^{p}$, (2) realized load $\overline{P}_{i,t}^{l,rt}$ and PV generation $\overline{P}_{i,t}^{pv,rt}$, (3) the SoC of BESS $E_{i,t}^{bess}$, (4) TSO signals: up-regulation price $\lambda_t^{\uparrow}$, down-regulation price $\lambda_t^{\downarrow}$, and the regulation state reg_t.

3) Action: The action of agent i at time t is:

$$a_{i,t} = \left[a_{i,t}^{pv}, \ a_{i,t}^{bess} \right] \in \mathcal{A}_i. \tag{12}$$

where $a_{i,t}^{pv} \in [0,1]$ scales the realized PV availability to set $P_{i,t}^{pv} \in [0, \overline{P}_{i,t}^{pv,rt}]$; and $a_{i,t}^{bess} \in [-1,1]$ scales the rated BESS power so that $P_{i,t}^{bess} \in [-\overline{P}_i^{bess,ch}, \overline{P}_i^{bess,dch}]$.

979-8-3315-8138-1

4) State Transition: The real-time dispatching uses a time step $\Delta t = 15$ minutes. At each step t, agent i observes its local information $o_{i,t}$ and selects an action $a_{i,t}$ according to its policy $\pi_i(o_{i,t}|a_{i,t})$. Let $\mathcal{I} = \{1, \ldots, I\}$ denote the set of agents and $a_t = \{a_{i,t}\}_{i\in\mathcal{I}}$ the joint action. The environment then evolves via the state-transition function $s_{t+1} = \mathcal{T}(s_t, a_t)$, where the next state s_{t+1} is determined by the current state and the joint actions of all agents.

5) Reward: The aggregator aims to: (i) maximize revenue from activated reserve capacity; (ii) reduce imbalance costs; (iii) minimize the penalties of reserve delivery deviations; (iv) exploit remaining flexibility for additional balancing revenue; and (v) reduce BESS degradation costs. Accordingly, a common reward is assigned to each agent i as follows:

$$\begin{aligned} r_{i,t} = &\left(R_t^{rt,\uparrow}\lambda_t^{B,\uparrow} + R_t^{rt,\downarrow}\lambda_t^{B,\downarrow}\right) + \left(E_t^{\uparrow}\lambda_t^{\uparrow} - E_t^{\downarrow}\lambda_t^{\downarrow}\right) \\ &+ \left(E_t^{sur} - E_t^{short}\right)\frac{\lambda_t^{\uparrow}g_t + \lambda_t^{\downarrow}h_t}{2} \\ &- \lambda_t^{B,-}\left(\Delta R_t^{\uparrow} + \Delta R_t^{\downarrow}\right) \\ &- \lambda_t^{deg}\,\Delta t \sum_{i\in\mathcal{I}}\left(P_{i,t}^{bess,c}\,\eta_i^{bess,c} + \frac{P_{i,t}^{bess,d}}{\eta_i^{bess,d}}\right) \end{aligned} \tag{13}$$

where $R_t^{rt,\uparrow/\downarrow}$ are the upward/downward reserve actually activated in real time, remunerated at $\lambda_t^{B,\uparrow/\downarrow}$. The terms $E_t^{\uparrow/\downarrow}$ denote additional upward/downward balancing energy if the aggregator still has available flexibility. E_t^{sur} and E_t^{short} represent surplus and shortage between day-ahead energy volume and real-time realizations, settled at the up-/down-regulation prices $\lambda_t^{\downarrow/\uparrow}$, depending on the system state (long or short) indicated by binary variables g_t and h_t. It should be noted that $E_t^{\uparrow}, E_t^{\downarrow}$ and $\Delta R_t^{\uparrow}, \Delta R_t^{\downarrow}$ represent the deviation surplus and shortage between the contracted reserve capacity and real-time provision. A surplus is rewarded in the same way as the balancing energy bids, while the shortage is penalized by a price $\lambda_t^{B,-}$, ensuring that the aggregator is incentivized to provide at least the required volume.

The energy balance between the day-ahead energy bid E_t^{da} and real-time realizations E_t^{rt} at the aggregator level is:

$$E_t^{rt} - E_t^{da} = \left(R_t^{rt,\uparrow} - R_t^{rt,\downarrow}\right) + \left(E_t^{\uparrow} - E_t^{\downarrow}\right) + \left(E_t^{sur} - E_t^{short}\right), \tag{14}$$

where

$$E_t^{rt} = \sum_{i\in\mathcal{I}} E_{i,t}^{p,rt} = \sum_{i\in\mathcal{I}} \Delta t\left(P_{i,t}^{bess} + P_{i,t}^{pv} - P_{i,t}^{l}\right) \tag{15}$$

6) Objective: Within an episode, each agent i collects a trajectory $\tau = \{o_{i,t}, a_{i,t}, r_{i,t}\}_{t=1}^{\mathcal{T}}$. The goal of each agent in the POMG is to learn a policy $\pi_i(a_{i,t}|o_{i,t})$ that maximizes the expected discounted return:

$$J_i(\pi_i) = \mathbb{E}\Big[\sum_{i\in\mathcal{I}} \gamma_t r_{i,t}\Big] \tag{16}$$

where $\gamma \in (0, 1]$ is the discount factor.

III. Methodology

A. Multi-agent deep deterministic policy gradient

MADDPG extends model-free DDPG to multi-agent settings under a CTDE framework [16]. Each agent $i \in \mathcal{I}$ has a deterministic actor $\mu_i(o_{i,t}; \phi_i)$ that maps its local observation $o_{i,t}$ to a continuous action $a_{i,t}$, and a centralized critic $Q_i(o_t, a_t; \theta_i)$ that evaluates the joint observation–action pair (o_t, a_t), where $o_t = (o_{1,t}, \ldots, o_{|\mathcal{I}|,t})$ and $a_t = (a_{1,t}, \ldots, a_{|\mathcal{I}|,t})$. Target networks μ_i' and Q_i' with parameters ϕ_i' and θ_i' softly track the online networks.

Experience replay and action selection. Joint transitions (o_t, a_t, r_t, o_{t+1}) are stored in a replay buffer $\mathcal{D}$, where $r_t = (r_{1,t}, \ldots, r_{|\mathcal{I}|,t})$. At time t, agent i selects its action $a_{i,t} = \mu_i(o_{i,t}; \phi_i)$ with added exploration noise $\epsilon_{i,t}$.

Critic update. From a mini-batch sampled from $\mathcal{D}$, the critic of agent i is updated by minimizing:

$$\mathcal{L}(\theta_i) = \frac{1}{B}\sum_{j=1}^{B}\Big(y_{i,j} - Q_i(o_t^j, a_t^j; \theta_i)\Big)^2, \tag{17}$$

where

$$y_{i,j} = r_t^{i,j} + \gamma\, Q_i'\big(o_{t+1}^j, a_{t+1}^j; \theta_i'\big), \tag{18}$$

where the target joint action $a_{k,t+1}^j$ is constructed from the target actors μ_k' for all agents.

Actor update. Given the centralized critic Q_i, the actor of agent i is updated using the deterministic policy gradient:

$$\nabla_{\phi_i} J(\phi_i) = \frac{1}{B}\sum_{j=1}^{B} \nabla_{\phi_i}\mu_i(o_t^{i,j}; \phi_i)\,\nabla_{a_i} Q_i(o_t^j, a_t^j; \theta_i)\big|_{a_i=\mu_i(o_t^{i,j}; \phi_i)}. \tag{19}$$

Target-network soft updates. After each gradient step, the target networks are updated via Polyak averaging:

$$\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i', \qquad \phi_i' \leftarrow \tau\phi_i + (1-\tau)\phi_i', \tag{20}$$

with $\tau \in (0, 1]$.

Training is centralized, while execution is fully decentralized. In deployment, each agent i uses only its learned actor μ_i^* to map local observations $o_{i,t}$ to actions $a_{i,t} = \mu_i^*(o_{i,t})$.

B. Attention-based MADDPG

To mitigate the curse of dimensionality in centralized critics with joint observation–action inputs, we adopt an attention-based extension of MADDPG (ATT-MADDPG). Instead of feeding the concatenated observations and actions of all agents into a multi-layer perceptron (MLP) critic, ATT-MADDPG employs a multi-head attention critic that selectively aggregates information from other agents, enabling the critic of agent i to focus on the most relevant counterparts. Under CTDE, the critic for agent i estimates $Q_i(o_t, a_t; \theta_i)$ using all agents' information and is parameterized as [17]:

$$Q_i(o_t, a_t; \theta_i) = f_i^Q\big(e_{i,t}, x_{i,t}\big), \tag{21}$$

where $f_i^Q(\cdot)$ is an MLP and $e_{i,t}$ is a learned embedding of agent i's own information,

$$e_{i,t} = f_i^E(o_{i,t}, a_{i,t}), \tag{22}$$

with $f_i^E(\cdot)$ a one-layer embedding network. The term $x_{i,t}$ aggregates the influence of the other agents via a multi-head self-attention mechanism:

$$x_{i,t} = W_i^o\ \mathrm{Concat}\big(\mathrm{head}_{i,t}^1, \ldots, \mathrm{head}_{i,t}^{N_H}\big), \tag{23}$$

where N_H is the number of attention heads and W_i^o is a projection matrix.

For each head $h \in \{1, \ldots, N_H\}$, embeddings $e_{j,t}$ are computed for all agents $j \in \mathcal{I}$ using the same embedding network, and queries, keys, and values are obtained by linear projections (e.g., $q_{i,t}^h = e_{i,t}W^{Q,h}$, $k_{j,t}^h = e_{j,t}W^{K,h}$,

979-8-3315-8138-1

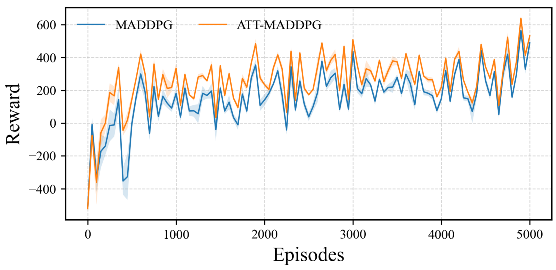

Fig. 2. Episodic training reward over 5,000 episodes.

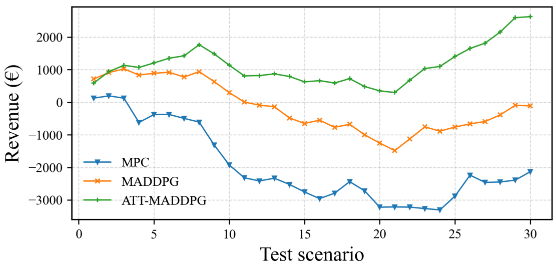

Fig. 3. Cumulative total reward over all test scenarios.

$v_{j,t}^h = e_{j,t} W^{V,h}$). The attention weight from agent i to agent j in head h is:

$$\alpha_{ij,t}^h = \frac{\exp\left(\frac{q_{i,t}^h (k_{j,t}^h)^\top}{\sqrt{d_h}}\right)}{\sum_{k \in \mathcal{I}\setminus\{i\}} \exp\left(\frac{q_{i,t}^h (k_{k,t}^h)^\top}{\sqrt{d_h}}\right)}, \tag{24}$$

where d_h is the head dimension, and the output of head h for agent i is:

$$\text{head}_{i,t}^h = \sum_{j \in \mathcal{I}\setminus\{i\}} \alpha_{ij,t}^h \, v_{j,t}^h. \tag{25}$$

By learning the attention weights jointly with the critic parameters θ_i, the model focuses on agents whose observations and actions are most informative for agent i's return while down-weighting less relevant interactions. ATT-MADDPG employs the same loss functions and target updates as MADDPG, but with gradients flowing through the attention mechanism, yielding a more expressive and scalable centralized critic for large, strongly coupled multi-agent systems.

IV. Case Study

A. Experiment setup

The proposed strategy is validated on an aggregator managing four prosumers that jointly participate in day-ahead energy and reserve markets and provide real-time balancing services. The experiments use real-world 15-minute smart meter data from Dutch residential prosumers, each equipped with a PV system and a BESS. Day-ahead electricity prices and upward/downward regulation prices for 2020 in the Netherlands are collected from ENTSO-E [18] and TenneT Transparency [19]. The data span one full year, with the first 11 months used for training and the last month for testing.

B. Training and testing performance

Figure 2 compares the training performance of ATT-MADDPG and vanilla MADDPG over 5,000 episodes. Both methods improve over time, indicating successful policy learning from exploratory to more coordinated behavior. However, ATT-MADDPG tends to converge to a higher reward with lower variance. This gain stems from its attention-based centralized critic, which learns to assign higher weights to the most influential agents when estimating each agent's Q-value, rather than treating all joint inputs uniformly as in vanilla MADDPG.

Figure 3 shows the cumulative total reward over the 30 test scenarios for the two trained MARL policies and an MPC baseline with a one-step prediction horizon. The one-step MPC acts as a greedy controller that optimizes only the current time step, while the MADDPG-based methods exploit long-term value estimates learned during training. ATT-MADDPG attains the highest cumulative reward, outperforming both MADDPG and the model-based baseline.

TABLE I

Economic results of single- and multi-service strategies over the one-month test period.

	Single-service strategy	Multi-service strategy
Day-ahead (k€)	18.75	29.82
Real-time (k€)	-8.4	2.63
Total (k€)	10.35	32.45

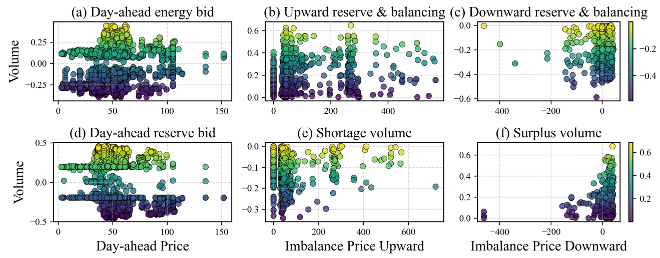

Fig. 4. Price-responsive allocation of energy, reserve, balancing, and imbalance volumes.

C. Optimal bidding strategy analysis

Deviations between day-ahead and real-time volumes arise from forecast errors and deliberate portfolio reallocation. The proposed ATT-MADDPG-based scheduler allocates flexibility to deliver contracted reserves, provide upward/downward services, and mitigate imbalances. Figure 4 shows the distribution of traded volumes versus prices. Day-ahead bids cluster around mid-range prices and are dominated by purchases, while reserve capacity forms a smaller share, with upward reserves clearly larger and more frequent than downward due to greater headroom for BESS discharging and PV curtailment (Fig. 4a–d). Consistently, real-time upward activations are more frequent and span a wider price range than downward activations (Fig. 4b–c). As actual activation depends on the TSO's signal, any shortage or surplus that is not contracted for any services is settled at imbalance prices. These volumes concentrate in low-price zones, limiting exposure to costly imbalance charges; additionally, and surplus occasionally yields profit when downward prices are non-negative (Fig. 4e–f). Overall, the distributions indicate a strategy that monetizes flexibility primarily through upward services, while steering residual imbalances into economically favorable price conditions.

Table I compares a single-service strategy with the multi-service provision strategy. In the single-service case, the aggregator trades only in the day-ahead energy market and

979-8-3315-8138-1

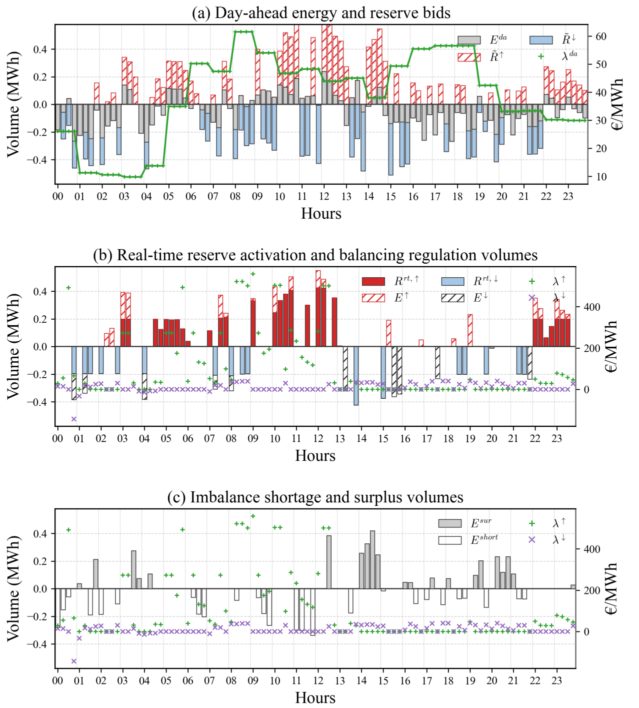

Fig. 5. Day-ahead and real-time dispatching results of the aggregator.

passively settles all deviations as imbalances, incurring high imbalance costs of 8.4 k€ and a lower overall profit. In contrast, the ATT-MADDPG-based multi-service strategy actively coordinates prosumer flexibility across energy, reserve, and balancing services, increasing day-ahead revenue to 29.82 k€ and achieving a real-time profit of 2.63 k€.

In Fig. 5(a), downward reserves are stacked on demand bids, exploiting the possibility to increase consumption or reduce generation when activated, while midday net demand bids are paired with upward reserves to enable PV curtailment and BESS discharging. Fig. 5(b) shows the reserves mobilized in real time by the TSO and the balancing regulation volumes delivered using the remaining flexibility. Upward regulation is pronounced in hours with high upward prices (around midday, 10:00–13:00), where BESS headroom is used to increase injections, while downward regulation is mainly activated in the early morning, afternoon, and evening hours when it is profitable to absorb surplus energy. In Fig. 5(c), surpluses are settled at downward prices and shortages at upward prices in a way that minimizes net imbalance costs. Thus, the MADRL-based strategy yields near-optimal real-time decisions that remain aligned with the co-optimized day-ahead reserve schedule, significantly enhancing the overall profitability of the aggregator.

V. Conclusion

Motivated by the uncertainty faced by aggregators on both the prosumer side (flexible demand) and the market side (price volatility), this study proposes a two-stage multi-service framework that combines day-ahead co-optimization with an ATT-MADDPG-based real-time dispatching strategy. Using real Dutch prosumer and market data, ATT-MADDPG achieves higher rewards than vanilla MADDPG and the MPC baseline across test scenarios, confirming the benefit of the attention-based centralized critic. The results show that ATT-MADDPG adapts well to uncertainty by orchestrating prosumer flexibility across reserve and balancing markets while reducing exposure to imbalance risks. Compared with the single-service (energy-only) strategy, the proposed multi-service strategy yields significantly higher overall aggregator profits over the one-month evaluation period. Future work will extend the framework to larger agent portfolios and explicitly incorporate network and congestion constraints.

Acknowledgment

The authors gratefully acknowledge Innova Energie for providing the datasets used in this study.

References

[1] C. Su, Y. Wang, and G. Strbac, "Coordinated electric vehicles dispatch for multi-service provisions: A comprehensive review of modelling and coordination approaches," *Renewable and Sustainable Energy Reviews*, vol. 223, p. 115977, 2025.

[2] J. P. Iria, F. J. Soares, and M. A. Matos, "Trading small prosumers flexibility in the energy and tertiary reserve markets," *IEEE Transactions on Smart Grid*, vol. 10, pp. 2371–2382, 2019.

[3] J. Iria, F. Soares, and M. Matos, "Optimal bidding strategy for an aggregator of prosumers in energy and secondary reserve markets," *Applied Energy*, vol. 238, pp. 1361–1372, 2019.

[4] J. Iria and F. Soares, "Real-time provision of multiple electricity market products by an aggregator of prosumers," *Applied Energy*, vol. 255, p. 113792, 2019.

[5] J. Wang, N. Xie, C. Huang, and Y. Wang, "Two-stage stochastic-robust model for the self-scheduling problem of an aggregator participating in energy and reserve markets," *Protection and Control of Modern Power Systems*, vol. 8, 2023.

[6] C. Loschan, D. Schwabeneder, G. Lettner, and H. Auer, "Flexibility potential of aggregated electric vehicle fleets to reduce transmission congestions and redispatch needs: A case study in austria," *International Journal of Electrical Power Energy Systems*, vol. 146, p. 108802, 2023.

[7] C. A. Correa-Florez, A. Michiorri, and G. Kariniotakis, "Optimal participation of residential aggregators in energy and local flexibility markets," *IEEE Transactions on Smart Grid*, vol. 11, pp. 1644–1656, 3 2020.

[8] C. O'Malley, L. Badesa, F. Teng, and G. Strbac, "Frequency response from aggregated v2g chargers with uncertain ev connections," *IEEE Transactions on Power Systems*, vol. 38, pp. 3543–3556, 2023.

[9] L. Wang, J. Kwon, N. Schulz, and Z. Zhou, "Evaluation of aggregated ev flexibility with tso-dso coordination," *IEEE Transactions on Sustainable Energy*, vol. 13, pp. 2304–2315, 2022.

[10] X. Wang, J. Wang, and J. Liu, "Vehicle to grid frequency regulation capacity optimal scheduling for battery swapping station using deep q-network," *IEEE Transactions on Industrial Informatics*, vol. 17, pp. 1342–1351, 2021.

[11] X. Sun and J. Qiu, "A customized voltage control strategy for electric vehicles in distribution networks with reinforcement learning method," *IEEE Transactions on Industrial Informatics*, vol. 17, pp. 6852–6863, 10 2021.

[12] Y. Tao, J. Qiu, S. Lai, X. Zhang, Y. Wang, and G. Wang, "A human-machine reinforcement learning method for cooperative energy management," *IEEE Transactions on Industrial Informatics*, vol. 18, pp. 2974–2985, 5 2022.

[13] Y. Wang, D. Qiu, G. Strbac, and Z. Gao, "Coordinated electric vehicle active and reactive power control for active distribution networks," *IEEE Transactions on Industrial Informatics*, vol. 19, pp. 1611–1622, 2023.

[14] D. Hu, Z. Ye, Y. Gao, Z. Ye, Y. Peng, and N. Yu, "Multi-agent deep reinforcement learning for voltage control with coordinated active and reactive power optimization," *IEEE Transactions on Smart Grid*, vol. 13, pp. 4873–4886, 2022.

[15] Y. Wang, D. Qiu, F. Teng, and G. Strbac, "Two-stage tso-dso services provision framework for electric vehicle coordination," *IEEE Transactions on Power Systems*, vol. 40, pp. 3117–3130, 2025.

[16] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in Neural Information Processing Systems (NeurIPS 2017)*, ser. NIPS'17. Curran Associates, Inc., 2017, pp. 6380–6391.

[17] S. Iqbal and F. Sha, "Actor-attention-critic for multi-agent reinforcement learning," pp. 2961–2970, 2019.

[18] European Network of Transmission System Operators for Electricity (ENTSO-E), "Entso-e transparency platform." [Online]. Available: https://transparency.entsoe.eu/

[19] Tennet, "Transparency data for the Dutch grid." [Online]. Available: https://www.tennet.eu/