Sustainable Energy, Grids and Networks 40 (2024) 101560

Contents lists available at ScienceDirect

Sustainable Energy, Grids and Networks

journal homepage: www.elsevier.com/locate/segan

Cooperative price-based demand response program for multiple aggregators based on multi-agent reinforcement learning and Shapley-value

Alejandro Fraija [a], Nilson Henao [a,*], Kodjo Agbossou [a], Sousso Kelouwani [b], Michaël Fournier [c]

[a] Department of Electrical and Computer Engineering, Hydrogen Research Institute, University of Québec at Trois-Rivières, Trois-Rivières, G8Z4M3, QC, Canada
[b] Department of Mechanical Engineering, Hydrogen Research Institute, University of Québec at Trois-Rivières, Trois-Rivières, G8Z4M3, QC, Canada
[c] Laboratoire des Technologies de l'Énergie (LTE), Centre de Recherche d'Hydro-Québec(CRHQ), Shawinigan, G9N7N5, QC, Canada

ARTICLE INFO

Keywords:
Demand response
Demand response aggregator
Dynamic pricing
Multi-agent reinforcement learning
Shapley-value

ABSTRACT

Demand response (DR) plays an essential role in power system management. To facilitate the implementation of these techniques, many aggregators have appeared in response as new mediating entities in the electricity market. These actors exploit the technologies to engage customers in DR programs, offering grid services like load scheduling. However, the growing number of aggregators has become a new challenge, making it difficult for utilities to manage the load scheduling problem. This paper presents a multi-agent reinforcement Learning (MARL) approach to a price-based DR program for multiple aggregators. A dynamic pricing scheme based on discounts is proposed to encourage residential customers to change their consumption patterns. This strategy is based on a cooperative framework for a set of DR Aggregators (DRAs). The DRAs take advantage of a reward offered by a Distribution System Operator (DSO) for performing a peak-shaving over the total system aggregated demand. Furthermore, a Shapley-Value-based reward sharing mechanism is implemented to fairly determine the individual contribution and calculate the individual reward for each DRA. Simulation results verify the merits of the proposed model for a multi-aggregator system, improving DRAs' pricing strategies considering the overall objectives of the system. Consumption peaks were managed by reducing the Peak-to-Average Ratio (PAR) by 15%, and the MARL mechanism's performance was improved in terms of reward function maximization and convergence time, the latter being reduced by 29%.

1. Introduction

The ever-growing demand for electricity and rapid electrification across economic sectors (leading to an increase in daily and seasonal energy peaks), combined with the problem of limited energy resources, awakens the importance of optimizing energy consumption. The immediate problem lies in traditional centralized approaches, which need to be enhanced to improve their ability to optimize energy demand and exploit the flexibility potentials of energy consumers. These centralized perspectives fall short of capturing the intricate dynamics of the complex and diverse power grid ecosystem and managing the evolving complexity of grid flexibility [1]. Consequently, the smart grid paradigm emerges, bringing with it the opportunity to facilitate the implementation of demand response (DR) programs, which are considered a viable option for managing energy demand by providing energy consumers a more active role [2]. These programs look for efficient solutions for minimizing generation costs, managing high demand peaks, reducing emissions, and improving the reliability of generation, transmission, and distribution systems [3]. They offer monetary incentives to induce changes in users' consumption patterns. The financial stimuli provide participants payments for reducing their consumption during periods of high demand or using time-varying price profiles to incentivize consumers to move their consumption to low-demand periods where lower prices are established [4].

In this context, a third-party entity is proposed called DR aggregator (DRA), which seeks to exploit the capacities of residential customers by implementing DR programs [5]. According to the literature, the role of DRAs is to group different agents in a power system to act as a single entity when participating in power system markets or selling services to the system operator. The management of users' flexibility potentials enables DRAs to participate on their behalf in the electricity market, where DRAs can identify flexibility potentials, automate their activation, and sell flexibility in electricity markets. Finally, DRAs can provide solutions to stabilize the revenues of market participants and bundle various services in the energy markets [6]. This, however, implies the need to determine monetary policies to maximize the DRAs' profit while offering a benefit to the users, leading the way to a new challenge [7]. For this reason, the policy generation problem has been

* Corresponding author.
E-mail address: nilson.henao@uqtr.ca (N. Henao).

https://doi.org/10.1016/j.segan.2024.101560
Received 9 April 2024; Received in revised form 29 September 2024; Accepted 1 November 2024
Available online 9 November 2024

Nomenclature

Acronyms

DR	Demand Response
DRA	Demand Response Aggregator
DSO	Distribution System Operator
EHS	Electric Heating System
IPPO	Independent Proximal Policy Optimization
MARL	Multi-Agent Reinforcement Learning
MG	Markov Game
PAR	Peak-to-Average Ratio
RL	Reinforcement Learning
SV	Shapley Value

Functions

$\lambda(\cdot)$	DSO reward in terms of PAR reduction
$\hat{A}^t$	Advantage at episode t
$\Lambda(\cdot)$	DRA welfare function
$\pi_k(\cdot)$	Price generator function
$\varphi^n(\cdot)$	Marginal contribution based on SV calculation
$f(\cdot)$	Thermal model
$PAR(\cdot)$	Peak-to-average ratio function
$R^{n,t}$	Reward function at episode t for nth DRA
$U(\cdot)$	Thermal comfort function
$v(\cdot)$	Characteristic function for coalition valuation
$Z(\cdot)$	Objective function for IPPO algorithm

Indices

j	House index
k	Time-step index
n	DRA index
t	Iteration index

Parameters

α	Rate of price change
λ^{max}	Maximum reward from DSO
π_0	Initial constant price
π_{min}	Lower price limit
M	Power value on inflexion point

Variables

$\boldsymbol{Y}^t$	System aggregated consumption at episode t
$\boldsymbol{y}^{n,t}$	Aggregated consumption at episode t for nth DRA
δ_k^j	Thermal discomfort factor of jth house
π_k^n	Price tariff defined by nth DRA at time-step k
$a^{n,t}$	Action at episode t for nth DRA
C	Coalition of DRAs
$o^{n,t}$	Individual observation at episode t for nth DRA
$s^{n,t}$	State at episode t for nth DRA based on system state and individual observation
S^t	System state at episode t
u_k^j	Energy consumption reported of jth house at time-step k
x_k^j	Indoor temperature of jth house at time-step k
x_k^{out}	Outdoor temperature at time-step k
x_{sp}^j	Set-point temperature profile of jth house
y_k	Aggregated energy consumption time-step k

addressed in the literature for different types of DR programs, from incentive-based to price-based [8]. Although the proposed approaches have made it possible to identify strategies for generating DRA's policies, as the number of aggregators increases, the challenge grows for utility companies to achieve load scheduling and produce reference signals for each of them [9].

In price-based DR programs, dynamic pricing has become one of the most influential and prominent strategies to encourage consumers to modify their consumption. However, defining an optimal policy to influence customers conveniently becomes challenging due to some uncertainties of load management. These uncertainties are related to the energy demand for each user, changing peak periods, and changes in the number of users and their preferences [10,11]. From the DRA perspective, there is also a need to propose policies guaranteeing aspects such as respect for user privacy throughout the strategy generation process [12]. This translates into increased uncertainty due to the significant lack of information in the decision-making process. As a result, reinforcement learning (RL) approaches have proven to be a valuable solution for dealing with the inherent uncertainties in different applications in DR context [13]. Nevertheless, when solving the price policy generation problem for a single DRA, it is not possible to guarantee that the individual solutions will lead to the best solution for the system. And, on the other hand, successfully implementing dynamic pricing with multiple DRAs requires a comprehensive evaluation and allocation of rewards among participating agents. This is where Shapley value (SV), a concept from cooperative game theory, comes into play [14].

SV is a classical mechanism from cooperative game theory, enabling the division of the total payoff so that each player receives a fair payment [15]. This method evaluates the marginal contribution of each player to the system and defines a uniquely equitable assignment of rewards, performing as a metric to measure the individual effort of each player [16]. As the main issue of the MARL mechanisms is that the actions performed by all agents influence the state transition, their interactions create a non-stationary environment from a single agent's view [17]. The proposed strategy demonstrates that combining SV to determine the DRAs' individual contribution alleviates the non-stationarity problem in the MARL-based multi-aggregator system, improving the obtained results during the training phase.

1.1. Related works

The definition of optimal dynamic pricing mechanism in DR programs is a relevant research topic that has been studied, and some solutions have been proposed. Its goal is to encourage users to change their consumption patterns to avoid generators' costly operation [18]. However, the definition of optimal price policies is a difficult task due to a lack of information on user preferences, price-responsive behavior linked to consumer flexibility, and the constantly changing energy load and energy generation of customers [19].

To address this problem, some authors have explored mechanisms to optimize the dynamic price policy generation decision-making process. For instance, the works done in [20] propose an optimization problem considering the stochasticity of renewable energy resources. In fact, the implementation of strategies where the objective function

of each player is embedded in one optimization problem is one of the approaches followed in the literature [21,22]. The problem with these approaches relies on affecting customers' privacy, negatively impacting user interest in participating in the DR program. To avoid this, authors have considered implementing game theoretical frameworks, in which the mechanisms seek to leverage their iterative process to reach an agreement and generate a price policy [23–25]. The problem is that the convergence process depends on the information customers provide. Therefore, this approach allows customers to cheat on the system to gain advantages, resulting in new challenges linked to the need to determine customers' trustworthy levels [26].

According to [27], RL techniques are well-known for their potential applicability in complex real-world applications, such as DR. In general, they are adaptable and capable of learning users' preferences through interaction without an explicit mathematical model. This makes RL an important tool for both sides, residential and supply, to properly define load control strategies and optimize price rates and incentives, respectively. For instance, authors in [28] utilized RL to optimize the objective function of the supply and demand side simultaneously. In regard to the aforementioned limitations, RL approaches have emerged as a valuable option to deal with problems related to the optimal price policy generation process. Authors, in [28], adopted a Q-learning method to decide the retail electricity price, considering service provider and customers profit, without requiring the full knowledge of the system dynamics and uncertainties. In [29], a deep Q network strategy was followed to build a dynamic subsidy price generation framework for a load aggregator avoiding the significant dependence on incorporating user feedback in its control loop. Furthermore, the works performed in [30,31] clearly stated the importance of RL in dealing with the lack of information about the customers' time-varying load demand and energy consumption patterns in the pricing optimization process. In addition to this, the authors in [32] even made use of this capability of RL algorithms to determine optimal pricing policies by learning from the price-responsive behavior of microgrids.

System operators may be unable to take on the additional effort of developing personalized price profiles for residents while determining their consumption patterns and preferences. This is due to the transaction costs and operational complexity that the system operator would otherwise have to bear when interacting with numerous individual buildings [33]. This is where DRA effectively facilitates customer participation by working in a more customer-oriented manner [34]. Particularly, multi-aggregator systems have only been addressed in a few works by implementing multi-agent systems. Authors in [9] implemented a hierarchical alternating direction method of multipliers (H-ADMM) to determine load following signals for multiple aggregators. In this mechanism, they assume aggregators have direct load control for individual devices, affecting customers' privacy and comfort. In [35], a bargaining-based cooperative game is proposed to solve irreconcilable incentive pricing strategies for multi-aggregators, where, again, the results depend on the excessive reliance on the users.

Considering RL approaches for determining dynamic pricing rates and multi-agent systems for multi-aggregator structures makes the MARL concept come into play. MARL has been gaining popularity in different smart grid scenarios, as it is presented in [36], due to its ability to deal with the inherent uncertainties of DR programs. These uncertainties can affect conventional approaches' performance, making them unsuitable for real-world implementations. In [37], active voltage control is proposed, based on Dec-POMDP, to enable real-world applications of MARL algorithms in power systems. Authors in [38] implemented a MARL approach to controlling a complex system of production resources, battery storage, electricity self-supply, and short-term market trading. In [39], authors demonstrate the value of MARL mechanism, which can quickly optimize thermostatically controlled load performance by applying collaborative multi-agent decision-making processes. In [40], an incentive-based DR program is considered based on MARL, which looks to maintain the capacity limits of the grid to prevent grid congestion by financially incentivizing residential consumers to reduce their energy consumption. In pricing strategies, authors in [41] proposed a mechanism design framework based on MARL to simultaneously determine the optimal charging prices for multiple charging stations over a period considering power output limits. In [42], authors developed a real-time pricing mechanism based on MARL where an RL-based grid agent defines a buy price to a set of RL-based prosumer agents. Finally, the works presented in [43], employ a cooperative-competitive MARL strategy based on Q-learning that enables the determination of optimal prices and incentives for maximizing benefits for both customers and service providers. This paper considers the effect of cooperation and competition among RL agents in the context of DR. However, these previous approaches have not considered fairness in the reward allocation process for each RL-based agent. For the specific case of DRAs, proposing MARL as a pricing approach for multi-aggregator systems makes determining a fair incentive allocation strategy necessary, as the definition of their rewards must be based on their individual contribution to the system operation performance. In [44], authors demonstrated that combining the DR programs with SV helps retailers assure profitability and also enhances user participation. Authors in [45,46] utilize SV to fairly divide the profit among microgrids and houses according to their efforts. These significant achievements presented in the literature highlight the potential of exploring the implementation of SV in a MARL-based multi-aggregator context for optimizing the exploitation of end-users' flexibility.

1.2. Motivation and contributions

This article delves deeper into dynamic pricing with multiple DRAs, where each DRA will determine price signals offering discounts based on customer responses in a cooperative game framework. The proposed mechanism incorporates a decentralized decision-making process, where each DRA aims to use its individual aggregated consumption profile as the only source of information to optimize the price policy generation process. However, for this purpose, it is necessary to face the uncertainties that appear in such a complex environment with incomplete information. Therefore, the implementation of an RL-based approach is proposed, that allows dealing with this type of scenario, in order to set the parameters of a dynamic price generator function. This enables the optimization of the tariff generation process, according to a global target set by the DSO. Accordingly, a mechanism based on MARL and SV-based reward-sharing mechanisms is described. The proposed cooperative MARL architecture harnesses the principles of game theory and RL to enable autonomous agents to learn and adapt to their environment. This approach ensures customers' privacy throughout the process of generating their optimal responses that minimize their costs and maximize their benefits. Each DRA will receive a reward from the Distribution System Operator (DSO) based on its individual contribution to peak shaving through the SV calculation. Integrating SV will provide a fair framework for distributing the benefits of cooperation among agents by assigning rewards to each agent's contribution and evaluating their marginal impact on the overall system. For brevity of the presentation, Table 1 compares the differences between the existing methods and the proposed model. Accordingly, this work contributes,

1. A cooperative price-based DR program for a set of DRA agents that cooperate to achieve better results in line with the DSO's objectives regarding peak shaving.
2. A cooperative MARL architecture to determine dynamic pricing strategies over the course of a coordination loop. The resulting price policies maximize the individual DRA's profit while providing gains to users.
3. A mechanism to fairly distribute the total gain of RL-based DRA agents through an SV-based reward-sharing mechanism. The calculation of its marginal contribution also speeds up the convergence process of the MARL algorithm.

Table 1
Comparison of state-of-the-art works.

Ref	DR Mechanism	Solution Method	Demand side target	Energy system	Objective function	Presence of DRA	Multiple aggregators	Price policy optimiza-tion	Reward sharing mechanism
[18]	Dynamic pricing	Price responsive modeling	Energy consumers	Deadline-Constrained electric loads	Estimation of consumer's price responsive behavior.	✗	✗	✓	✗
[19]	Dynamic pricing	Bi-level, meta-heuristic	Customers with smart meters	Interruptible, non-interruptible and curtailable appliances	Maximize retailer profit while customers aim to minimize their electricity bills.	✗	✗	✓	✗
[20]	Real-time pricing	Stochastic optimization	Residential, commercial and industrial customers	Energy storage systems	Minimize the cost for end-user customers and increasing the retailer profit while flattering the load profile.	✗	✗	✓	✗
[21]	Dynamic pricing	Multi-objective optimization	Energy demanders in micro-grids	electricity, heat and cool loads	optimal value of energy prices under different entities interest.	✗	✗	✓	✗
[22]	Time-of-Use	Multi-objective optimization	Residential consumers	Shiftable appliances	Maximize consumer surplus by adjusting the electricity price, guaranteeing a fixed profit to the utility company.	✗	✗	✓	✗
[23]	Dynamic pricing	Game-theoretic model	Residential consumers with thermostatically controlled loads	Thermostatically controlled loads	Maximize the social welfare defined as the sum of consumer surplus and retail profit.	✗	✗	✓	✗
[24]	Time-of-Use	Game-theoretic model	DRAs	Schedulable loads	Minimize the total cost of purchasing electricity from the bulk market for the utility company and maximize DRAs payoff function.	✓	✗	✓	✗
[25]	Time-of-Use	Game-theoretic model	Residential customers	Heterogeneous loads	Minimize the player's costs based on the predicted strategy of all other players.	✗	✗	✓	✗
[28]	Dynamic pricing	RL	Residential customers	Curtailable loads	Maximize service provider's profit and minimize customers' costs.	✗	✗	✓	✗
[29]	Dynamic pricing	RL	Customers contracted with a wind farm	Electric heating system	Maximize the load aggregator revenue.	✗	✗	✓	✗
[9]	Load following signals	H-ADMM	Customers with controllable HVAC systems	HVAC systems	Minimize the penalty for drawing power beyond a predefined limit, and minimize customer discomfort.	✗	✓	✗	✗
[30]	Real-Time pricing	RL	Residential customers	Mathematical response function	Minimize the total expected system cost.	✗	✓	✗	✗
[31]	Dynamic pricing	RL	Residential customers	Accumulated load demand	Minimize the expected total cost or customers' disutility.	✗	✓	✗	✗
[32]	DLMP	RL	Microgrids	Dispatchable generator	Maximize the profit from selling energy while minimizing the PAR.	✗	✓	✗	✗
[35]	Diverse compensation price	Game-theoretic model	Energy consumers	Curtailable loads	Minimize the electric utility cost of the electric utility company, minimize DRA cost and maximize its revenue, and maximize customer incentive and minimize its discomfort.	✓	✓	✓	✗
[40]	Incentive-based	MARL	Residential consumers	Curtailable and shiftable appliances	Minimize financial costs for the aggregator while maintaining the capacity limits of the electricity grid and preventing grid congestion.	✗	✗	✗	✗
[41]	Dynamic pricing	MARL	Charging stations	Electric vehicle	Maximize the long-term network revenue considering the social welfare of all users.	✗	✗	✓	✗
[42]	Dynamic pricing	MARL	Prosumers in micro-grids	Energy storage and PV systems	Maximize the long term profit of players.	✗	✗	✓	✗
[43]	Real-time pricing, Time-of-Use, direct load control	MARL	Customers with elastic loads	Elastic loads	Maximize benefits for both customers and electric Service Provider	✗	✗	✓	✗
Proposed work	Dynamic pricing	MARL	Residential customers	Electric heating system	Maximize DRAs profit while reducing the consumption peaks of the system.	✓	✓	✓	✓

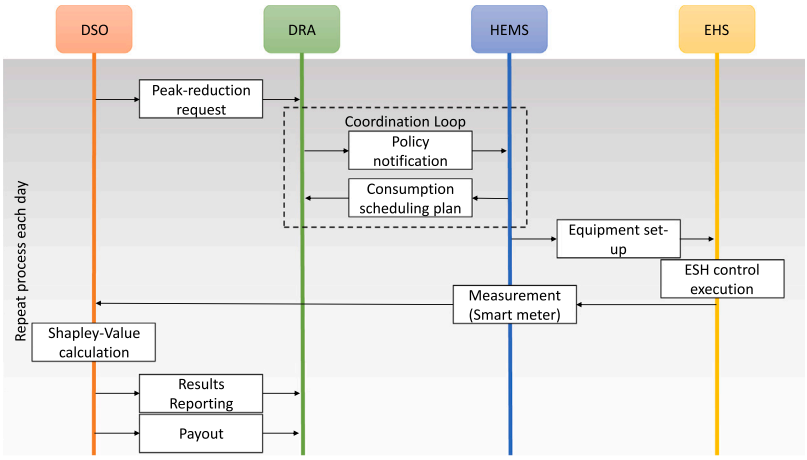


Fig. 1. Automatic DR sequence for the multi-aggregator system.

The rest of the paper is organized as follows: Section 2 summarizes the methodology for the developed MARL framework. The case study is discussed in Section 3, followed by the conclusion in Section 4.

2. DR mechanism and problem formulation

DSOs are expected to explore the distribution-level flexibility potential for tackling grid problems, making reducing the system's peak power one of its goals. For this reason, the DSO interacts with a group of DRA agents who will manage the flexibility of different groups of houses. As presented in Fig. 1, the DSO rewards each DRA for contributing to the peak shaving objective in a day-ahead scheme. In response, the DRA stipulates price policies through a coordination loop, where the DRA acts as a leader of the group of residential agents that respond with a consumption plan until an agreement is reached. The dynamic price policies based on discounts induce customers to modify their consumption patterns, while the DRA performs a trade-off between the profit of selling energy to residential customers and the DSO's monetary incentive for peak shaving. The coordination loop is performed at the beginning of the day, and once the agreement is reached, the price profile is established, and residential customers are committed to following their consumption plans during the day according to the contract defined with the DRA. At the end of the day, DSO verifies the improvement of the consumption demand and the contribution of each DRA by means of the SV-based reward-sharing mechanism to determine their rewards. Fig. 2, provides a representation of the interaction between the different actors of the proposed scenario. As seen in Figs. 1 and 2, the only information that each DRA uses to define the pricing policy is the consumption profile reported by each customer. This guarantees respect for users' privacy but generates a high complexity in the policy optimization process due to the lack of information. It is for this reason that a MARL approach is proposed below. Finally, Even though there is no information exchange between the different DRAs, there exists an interdependence between them, as the action performed by each aggregator significantly impacts the performance or behavior of others, due to their individual contributions to the collective goal, ending in the need to cooperate [47].

2.1. DRA agents

From the upper level, the DRAs communicate their aggregated consumption plans to the DSO before implementing a dynamic pricing mechanism, i.e., with a constant price π_0. It is assumed that all players communicate truthful information in this first interaction since the analysis of the effect of perverse players is out of the scope of this work. With this information, the DSO establishes a reward λ for the DRAs that depends on the peak shaving of the load profile. For this, the DSO utilizes the peak-to-average ratio (PAR), which is used to measure the effectiveness of the demand-side management algorithms [48]. The DSO considers the overall PAR ratio as a mechanism to determine the reduction of the overall peak demand. Dividing a one-day period in K timestamps, the calculation of this ratio is performed over the total aggregated load demand $\boldsymbol{Y} = \{Y_1, \ldots, Y_K\}$, as follows:

$$PAR(\boldsymbol{Y}) = \frac{\max_k\{\boldsymbol{Y}\}}{\frac{1}{K}\sum_{k=1}^{K} Y_k} \tag{1}$$

At the bottom level, each DRA interacts with its group of residential agents as retailers in a Stackelberg game. As a leader, each DRA seeks to optimize its profits that depend on its individual income from selling the energy to the set of customers. However, in order to gain the advantage of the reward offered by the DSO, the DRA defines discounts during the day to incentivize users to change their consumption patterns. The utilization of these discounts will guarantee a reduction of the customers' bills, with respect to their normal consumption when an initial constant price π_0 is established. In this way, each DRA benefits from the coordination loop, using the aggregated consumption plan of the houses $\boldsymbol{y} = \{y_1, \ldots, y_K\}$ as the only source of information as a privacy-preserving approach. To ensure the generation of price profiles considering the upper limit as the constant price π_0 and the lower limit linked to the least price value to be offered by each DRA, the aggregator applies a monotonic transformation of $\boldsymbol{y}$ based on the logistic function to determine $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ as follows:

$$\pi_k(y_k) = \pi_{min} + \frac{\pi_0 - \pi_{min}}{1 + \exp\left(\frac{-y_k + M}{\alpha}\right)} \tag{2}$$

where π_{min} is the minimum price that each DRA is willing to offer to its customers, α is a parameter to control the price rate of change, and M is the power value where the inflection point of the function is set. According to [49], this function provides a better approach for exploiting the flexibility potentials from the residential sector in a more controllable way, when it is utilized in a coordination loop with a regularization of the residential agents' response. Translating the M value as the target for maximum power consumption of the daily profile. The monotonic transformation will allow as well the parameterization of the pricing policy to reduce the complexity of calculation in its generation, ensuring the generation of higher price values when consumption is higher and lower price values during lower consumption periods. Once the new price profile is generated, it is communicated to the customers, which will replay with a new plan until an agreement is reached. Therefore, the benefit of each DRA can be explained by the trade-off between the profit from selling the

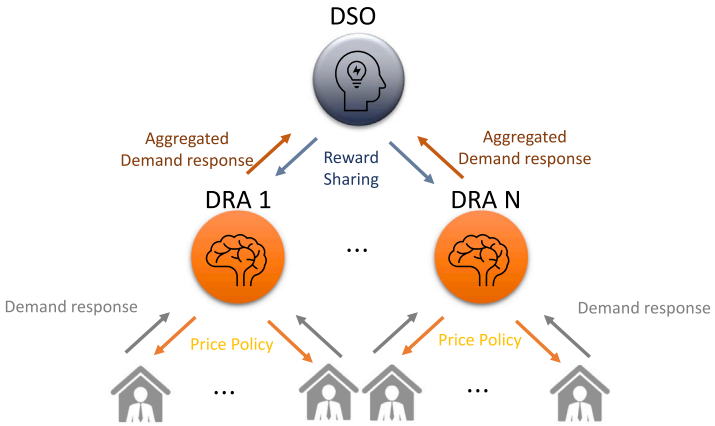


Fig. 2. Interaction between market participants in the DR program.

energy to its customers and the reward received from the DSO from contributing to the peak shaving objective,

$$\arg\max_{\boldsymbol{\pi}} \Lambda(\boldsymbol{\pi}) = \omega_1 \lambda[PAR(\boldsymbol{Y})] + \omega_2 \sum_{k=1}^{K} \pi_k y_k \tag{3}$$

where ω_1 and ω_2 are weighting factors to balance these two terms, and $\lambda(\cdot)$ is the DSO's function to calculate the reward in terms of PAR. These ω values allow the assignation of a lighter or heavier importance to each term of the objective function. To properly establish these parameters, several simulations are performed to obtain a dataset of different results for each term of the reward function. Finally, each of these weighting factors is defined first by the inverse of the unweighted average of each term to guarantee a normalized result; after that, these values can be slightly modified if it is necessary to give more importance to any of the terms in (3). As the proposed approach does not consider a convex PAR-related metric, this objective function cannot be treated with the classical gradient-based optimization approaches, as the PAR function itself of the total aggregated system consumption is not convex. Moreover, as the reward λ depends on the aggregated performance of the DRAs, it is necessary to determine a fairness strategy to determine the reward for each aggregator in terms of its marginal contribution. Consequently, the MARL architecture is implemented to deal with the intractability of the DRAs' objective function for optimizing the dynamic pricing decision-making process. Furthermore, an SV calculation is implemented to determine the marginal contribution of each DRA in the proposed scenario.

2.2. Cooperative MARL method for multi-aggregator system

Overview of MARL. RL algorithms are machine learning techniques based on a trial-and-error process for sequential decision-making problems. In a single-agent RL mechanism, an agent interacts with an unknown environment by executing actions to extract useful information, and the environment responds with an immediate reward to evaluate the selected action. The agent aims to maximize its reward by realizing a trade-off between exploring new actions and exploiting those who seem optimal. This strategy is advantageous in scenarios such as the one proposed in this paper, where DRAs need to determine a price policy relying only on the information of their daily consumption plan. The absence of relevant information, combined with the high-dimensional, non-convex nature of the problem, and the lack of a predefined price-responsive mathematical model, pose significant challenges for classical optimization methods. In contrast, RL offers a distinct advantage in managing these complex decision-making scenarios by effectively navigating uncertainty and non-linearity. However, a sacrifice needs to be made in order to obtain the information needed to optimize the price policy generation process, which is related to the agents' learning period and convergence guarantees. RL often requires large amounts of data and significant computational resources for training. Furthermore, classical optimization algorithms, especially for convex problems, have well-established convergence guarantees. RL algorithms might converge slowly or even fail to converge in complex, non-stationary environments.

Moving to MARL, new relationships appear between agents in the same environment that compete or cooperate between them to maximize their rewards. As a result, agents' rewards are influenced by states and actions performed by the other RL agents. Mathematically speaking, in single-agent RL approaches, the interactions between the environment and the agent are modeled by a Markov Decision Process (MDP). In the case of MARL, these interactions are based on a Markov game (MG), a combination of MDP and game theory [50]. For these reasons, this work proposed the combination of the MARL architecture with an SV-based reward-sharing mechanism. This approach mitigates the cross-influence between RL agents while enhancing model convergence.

Markov game formulation. The proposed scenario considers a multi-agent system composed of RL-based DRAs, each interacting with their own residential customer group. To explore the generation of dynamic pricing strategies, the interactions between the residential agents and the RL agents are modeled by a finite MG. Therefore, the components required are: N agents corresponding to N DRAs. A shared state set $\mathscr{S}$ and the collection of agents' private observation sets $\{\mathscr{O}_{1,\dots,N}\}$. The action sets $\{\mathscr{A}_{1,\dots,N}\}$ and individual reward sets $\{\mathscr{R}_{1,\dots,N}\}$. And a set of state transition functions $\{\mathscr{P}_{1,\dots,N}\}$. Considering the state 0 of the system as the aggregation of the users' consumption plan when all the DRAs establish a constant price. The proposed scenario defines an episode for the MARL mechanism as the coordination loop between DRAs and residential agents, where each step comprises the definition of a price signal from the DRAs with its associated DR. The MG components are stated as follows:

1. *System state and MG observations*: The system state S^t is described by the aggregated power consumption profile of the system $\boldsymbol{Y}$ normalized concerning the maximum power consumption $\max_k \boldsymbol{Y}^0$ presented in the consumption plan of the user when initial constant prices are established. Similarly, the individual private observation for agent n is defined as $o^{n,t}$, described by the aggregated power consumption profile of its customers $\boldsymbol{y}^n$ normalized to the maximum initial power consumption $\max_k \boldsymbol{y}^{n,0}$.

2. *MG Actions*: For each agent n the action $a^{n,t} = \{M, \alpha, \pi_{min}\}$ modifies its price generator function presented in Eq. (2), where M values can go from the initial aggregated average consumption $\frac{1}{K}\sum_{k=1}^{K} y_k^0$ to the maximum consumption $\max_k\{\boldsymbol{y}^0\}$.
3. *Reward functions*: Finally the reward function for the n agent is $R^{n,t}$.

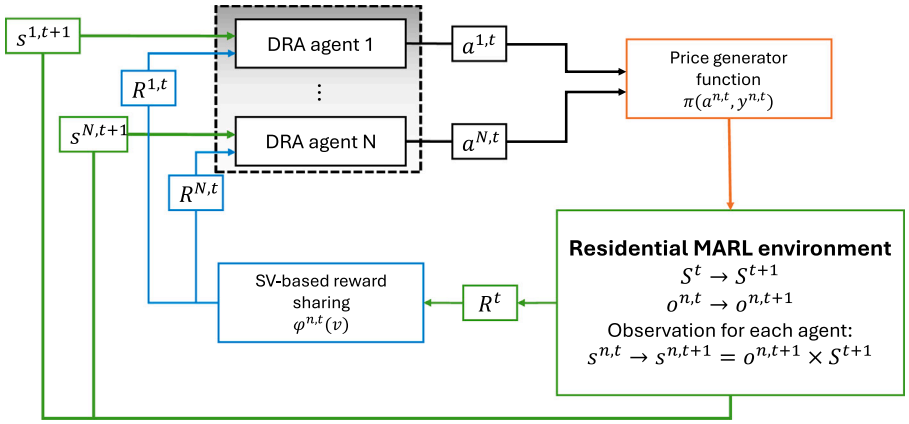


Fig. 3. Representation of the proposed Markov Game.

To avoid an improper calculation of rewards for each DRA, it is necessary to utilize a fair strategy to calculate the individual contribution of each DRA on the system peak shaving. This strategy will modify the reward functions of the MG, improving the agents' understanding of the impact of their actions on the environment [51]. To better understand the interaction between the residential agents and the RL-based DRA agents, Fig. 3 provides an illustrative explanation of the Markov Game for the proposed methodology. The explanation of the fair strategy based on SV and the final agents' reward function is explained below.

Shapley-Value based reward sharing mechanism. The DSO seeks to determine rewards fairly for the DRAs, according to the objective established by him, and the marginal contribution of each DRA. For this purpose, a total reward function is defined to determine the total reward that DSO will distribute between DRAs. This reward function is inversely proportional to the PAR of the system aggregated load profile. The utilized function $\lambda(\cdot)$ is based on the same proposed by [46], as follows:

$$\lambda[PAR(\boldsymbol{Y})] = \frac{1}{1 + e^{c_1(PAR(\boldsymbol{Y}) - c_2)}} \lambda^{max} \tag{4}$$

c_1 and c_2 are function parameters defined by the DSO to adjust the reward function shape, and λ^{max} is the maximum reward for PAR reduction. The c_1 and c_2 values are preset after a negotiation process between the system operator and DSO. This ensures the definition of the reward function before starting the learning process of the RL agents. Such an idea follows the methodology presented in [46]. Furthermore, this work also proposes the values for these parameters for a load factor-based reward function. Thus, the choice of c_1 and c_2 will be based on the equivalent PAR-based representation. Finally, the reward λ^{max} is based on a proportion of the operational and generation cost reduction.

By creating a grand coalition, the DRAs collaborate looking for maximizing individual and system objectives. As the contribution of each player might be different, it is necessary to measure each DRA's contribution to the peak shaving achievement for determining the allocation of the total payoff. With N DRAs and a function v that maps subsets of DRAs to the real numbers. The amount that a DRA n receives in the given coalitional $(v, \mathbb{C})$ game is,

$$\varphi^n(v) = \sum_{C \subseteq \mathbb{C} \setminus \{n\}} \frac{|C|!(N - |C| - 1)!}{N!} (v(C \cup n) - v(C)) \tag{5}$$

where $\mathbb{C}$ represents the set of all possible coalitions, C is a subset of $\mathbb{C}$, $|\cdot|$ determines the cardinality of the given set, and $v(C)$ represents the valuation for the coalition C. The sums is done over all coalition subsets not containing the DRA n. The contribution of each DRA n is calculated for all C based on the expression $v(C \cup n) - v(C)$, and then the average of these contributions is calculated to determine the fair allocation of its reward. Finally, the characteristic function is designed as:

$$v(C) = \frac{\|\boldsymbol{y}^{C,0} - \boldsymbol{y}^{C,t}\|_2^2}{\|\boldsymbol{Y}^0 - \boldsymbol{Y}^t\|_2^2} \tag{6}$$

$y^{C,0}$ represents the aggregated profile for the coalition C in state 0, i.e, for the constant price, and $y^{C,t}$ is the aggregated profile after the implementation of the dynamic pricing mechanism. Likewise, Y^0 and Y^t present the aggregated profiles of the system.

Independent Proximal policy optimization (IPPO) method.

Algorithm 1: IPPO algorithm

For each DRA agent n:
DRA asks residential agents for their stipulated consumption plan under the initial constant price π_0, and defines $o^{n,0}$.
DRA communicates the aggregated plan to the DSO, which returns the system aggregated profile state S^0 for defining the initial state $s^{n,0} = \{S^0\} \times \{o^{n,0}\}$
for $t = 0, 1, 2, \ldots$ **do**
  Define the action $a^{n,t} = \{M^n, \alpha^n, \pi_{min}\}$. (*Price function transformation defined by DRA n*).
  Calculate the pricing profiles based on (2) using $a^{n,t}$ and send them to the residential agents.
  Residential agents solve their optimization problems according to (13)
  DRA communicates to the DSO its aggregated consumption plan and defines its individual observation $o^{n,t}$.
  DSO calculates its individual contribution $\varphi^{n,t}(v)$ with Shapley-Value, based on equations (5) and (6).
  DSO communicates the reward calculated based on (4), and the system aggregated profile S^t.
  Get the normalized state $s^{n,t}\{S^t\} \times \{o^{n,t}\}$. (*cartesian product between the system state and its individual observation*).
  Calculate rewards $R^{n,t}$.
  Collect the set of partial trajectories $\{(s^{n,t}, a^{n,t}, R^{n,t}, s^{n,t+1})\}$ on policy $\phi^{n,t} = \phi_{\theta^{n,t}}(a^{n,t}, s^{n,t})$.
  Estimate advantage $\hat{A}^{n,t}$.
  if t *mod* $T = 0$ **then**
    Compute policy update

$$\theta^{n,t+1} = \arg\max_{\theta} \sum_{j=0}^{T} Z(\theta)$$

    via stochastic gradient ascent with Adam [52].
  end
end

As the customers are different for each DRA, the actions needed during each coordination process are different. It means that each RL-based DRA must learn its own best strategies independently. For this purpose, an Independent Proximal Policy Optimization (IPPO) technique is proposed. According to [53], empirical studies have shown that IPPO can offer excellent performances, close to or even better than the MARL techniques based on centralized training with decentralized execution, in several benchmarks. This algorithm is a cooperative MARL strategy where each RL agent learns independently using PPO. PPO is a practical and effective policy gradient algorithm derived from Trust Region Policy Optimization (TRPO), that replaces a trust region constraint with a simpler clip trick. The algorithm uses a parameter θ to optimize a policy $\phi_\theta(a^t, o^t)$. In RL theory, this policy describes the agent's behavior in deciding the action that must be performed in a given state. Using the clip trick, this technique stabilizes the training process by avoiding high policy alterations during the parameter updating process. This trick attempts to keep old and new policies closer, resulting in reward enhancement and stability [54]. The parameter updating of θ is achieved by maximizing the objective function,

$$Z(\theta) = \hat{\mathbb{E}}^t[\min(r^t(\theta)\hat{A}^t, clip(r^t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}^t)] \tag{7}$$

where $\hat{E}^t$ is the expectation over episode t, $r^t(\theta)$ presents the probability ratio between the new and old policies in terms of $\phi_\theta(a^t, s^t)/\phi_{\theta_{old}}(a^t, s^t)$. ϵ is the hyperparameter for clipping to avoid large deviations in the θ updating process. And $\hat{A}^t$ is the advantage estimation to measure the performance of the selected action given the current state, using the RL value function $V(s^t)$, the discount factor γ and the batch size T, and is calculated as follows:

$$\hat{A}^t = -V(s^t) + \gamma R^t + \cdots + \gamma^{T-t+1} R^{T-1} + \gamma^{T-t} V(s^T) \tag{8}$$

s^t and R^t are the state and the reward on episode t for each RL agent, respectively. Being the system state S^t the only shared information between the DRA agents, for the proposed scenario, the state $s^{n,t}$ for the DRA n will be established as the Cartesian product $S^t \times o^{n,t}$ between the system state and its individual observation, i.e., $s^{n,t} = \{S^t, o^{n,t}\}$. Furthermore, combining the Eqs. (3) and (5), the individual reward at state s^t for agent n can be finally stated as follows:

$$R^{n,t} = \omega_1 \varphi^{n,t}(v)\lambda[PAR(\boldsymbol{Y}^t)] + \omega_2 \sum_{k=1}^{N} \pi_k^n(a^{n,t}) y_k^{n,t} \tag{9}$$

The Algorithm 1 represents the utilized IPPO technique, based on the PPO mechanism presented by [52].

2.3. Automated DR for residential agents

For the case of the residential agent, it is assumed that each of them is equipped with a home energy management system (HEMS). The HEMS deals with controllable and non-controllable loads to modify the consumption plan by scheduling the consumption of the flexible ones. In this case, the controllable load refers to electric heating systems (EHS) controlled by smart thermostats, and the non-controllable loads are the other household appliances. Based on end-users comfort, the HEMS can modify the heating consumption to provide the flexibility required for residential agents to gain an advantage from the discounts offered by the dynamic pricing mechanism. Subsequently, the individual welfare maximization for each user j, can be expressed by,

$$\begin{aligned} \underset{\mathbf{u}^j=\{u_k^j\}_{k=1}^K}{\text{Maximize}} \quad & \sum_{k=1}^{K} (U(u_{h,k}^j) - \pi_k^n u_k^j) \\ \text{subject to} \quad & x_{k+1}^j = f(x_k^j, x_k^{\text{out}}, u_{h,k}^j, \boldsymbol{w}^j) \\ & x_k^j \in [x_{\min}^j, x_{\max}^j] \\ & u_k^j \in [0, u_{\max}^j] \\ & u_k^j = u_{h,k}^j + u_{fix,k}^j \end{aligned} \tag{10}$$

where the vector $\boldsymbol{u}^j = \{u_1^j, \ldots, u_K^j\}$ represents the consumption plan of the jth house, considering the aggregation of thermal and fixed loads, $u_k^j = u_{h,k}^j + u_{fix,k}^j$. As the residential agent interacts with the DRA n, π_k^n is the dynamic tariff this aggregator defines at timestamp k. The parameters $x_{\min}^j$ and $x_{\max}^j$ are the lower and upper bounds for the allowed internal temperature according to users thermal preferences, respectively, and $u_{\max}^j$ is the maximum heating system capacity in time slot k. $f(\cdot)$ is a linear model for describing the dynamic thermal response of the house. This model depends on the indoor temperature x_k^j, the outdoor temperature x_k^{out}, the heating power consumption $u_{h,k}^j$ and the matrix coefficients $\boldsymbol{w}^j$. According to [55,56] this model can be expressed as:

$$\begin{aligned} x_{k+1}^j &= f(x_k^j, x_k^{\text{out}}, u_{h,k}^j, \boldsymbol{w}^j) \\ &= w_1^j x_k^j + w_2^j x_k^{\text{out}} + w_3^j u_{h,k}^j. \end{aligned} \tag{11}$$

The first term in Eq. (10) refers to the customer's utility function; the second term is the customer's cost expressed by the bill to pay. The utility function $U(u_k^j)$ models the thermal user's thermal comfort and is determined by the set-point temperature x_{sp}^j and δ_k^j, the comfort weight factor representing the user's elasticity. δ_k^j explains how much users are willing to sacrifice their comfort to reduce the bill, and it is also used for weighting the utility with respect to the cost [55]. This comfort factor is a daily profile based on a historical analysis of set-point profiles. This means that the user's elasticity changes in time during the day. The δ_k^j can take values from the set $[0, \delta_{\max}^j]$, following the set-point shape profile, assuming that higher values of set-points mean higher thermal comfort needs. In the case of $\delta_k^j = \delta_{\max}^j$, occupants are inelastic, and they are interested in maintaining their comfortable temperature set-point. Conversely, the agent can freely modify the internal temperature when $\delta_k = 0$ while respecting the constrain $x_k^j \in [x_{\min}^j, x_{\max}^j]$. This strategy maximizes the flexibility potentials of the residential agent while respecting its thermal comfort constraints. Finally, according to [57], the residential thermal comfort function can be modeled through,

$$U(u_{h,k}^j) = -\delta_k^j (x_{\text{sp}}^j - x_k^j)^2, \tag{12}$$

The residential agents receive the price policy from the DRA simultaneously and selfishly solve their optimization problems. In order to make them coordinate through the coordination loop, it is necessary to regularize their decision-making process. The proposed regularization strategy is based on proximal decomposition as a distributed algorithm [58]. For this, a regularization parameter, τ, is utilized to penalize differences between consecutive defined consumption plans through the coordination loop, i.e., penalize significant variations between episodes t and $t-1$ [59]. Thus, the dual optimization problem residential agents' cost function can be defined by (13).

$$\begin{aligned} \underset{\mathbf{u}^j=\{u_k^j\}_{k=1}^K}{\text{Minimize}} \quad & \sum_{k=1}^{K} \delta_k^j (x_{\text{sp}}^j - x_k^j)^2 + \pi_k^n u_k^j + \tau (u_k^{j,t} - u_k^{j,t-1})^2 \\ \text{subject to} \quad & x_{k+1}^j = f(x_k^j, x_k^{\text{out}}, u_{h,k}^j, \boldsymbol{w}^j) \\ & x_k^j \in [x_{\min}^j, x_{\max}^j] \\ & u_k^j \in [0, u_{\max}^j] \\ & u_k^j = u_{h,k}^j + u_{fix,k}^j \end{aligned} \tag{13}$$

For each residential group interacting with a DRA, it is necessary to determine the regularization parameter that ensures the convergence of the mechanism and the avoidance of rebound peaks in the day-ahead market [60]. The selection of the τ value is performed according to the inequality:

$$\tau > 4(J_n - 1)\pi_0 \tag{14}$$

where J_n is the number of residential agents interacting with the DRA n and π_0 is the initial constant price [56]. For each iteration,

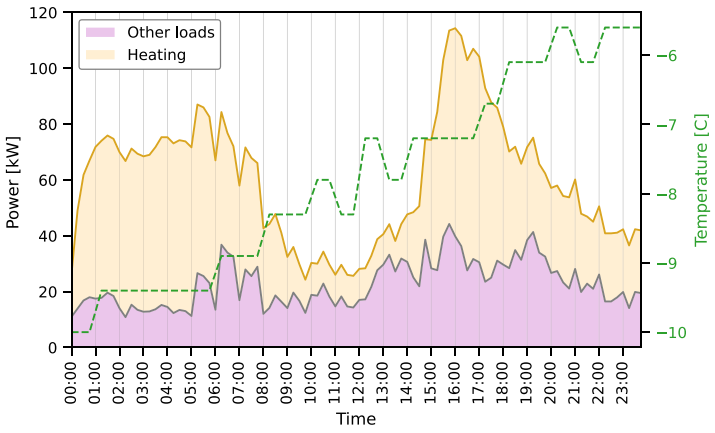


Fig. 4. Aggregate energy demand when exposed to a winter outdoor temperature profile.

the DRA sends the new price policy, and the residential agent solves the optimization problem (13) until an agreement is reached. Finally, the coordination process between the DRA and the residential agents converges when the relative PAR of the aggregated profile between successive iterations is lower than a predefined threshold:

$$\frac{\|PAR(\boldsymbol{y}^t)\|_2}{\|PAR(\boldsymbol{y}^{t-1})\|_2} < threshold \tag{15}$$

3. Results and discussion

This section provides the simulation results of the proposed MARL-based DR mechanism. First, a validation of the residential consumption behavior model is carried out. Then, the training process results are examined through the learning process of the best parameters selection for the price function during the coordination loop and the results in peak-shaving of the IPPO-based RL technique combined with the SV-based reward-sharing mechanism. Finally, the importance of the SV is presented, and how it improves the performance of the proposed MARL technique.

Residential agents behavior. The system environment for validating the proposed technique comprises 11 residential agents. We collected data from 11 single-family detached houses in Trois-Rivieres, Quebec, Canada, during a winter period (from January to April 2018), with a 15-minute sampling interval. The houses are equipped with electrical baseboards and controllable thermostats for temperature control. Using the real-world data, we constructed the thermal models for all the residential agents, considering the recorded indoor temperatures, the electrical heating power consumption, and the outdoor temperature. And a ridge regression mechanism was applied to determine the matrix coefficients $\boldsymbol{w}^j$ needed in Eq. (11). Furthermore, statistical information from a previous study conducted in [61] is utilized to randomly generate the set-point values x^j_{sp} from the set $\{20, 21, 22, 23\}$ in degree Celsius [C]. The different levels of users' thermal elasticity δ^j_k for the utility functions can be extracted from a log-normal distribution with the expectation, $\mathbb{E}(\delta_{\max}) = 5$, and variance, $Var(\delta_{\max}) = 1$. Finally, with the historical power consumption of energy-extensive appliances other than electric boards, an aggregate load profile of non-controllable loads is generated and added to the simulated heating consumption.

Fig. 4 shows the aggregated consumption behavior of the residential users exposed to a temperature profile of a winter day. The behavior shown in the Figure demonstrates that the developed residential models follow the expected power consumption pattern of Quebec's residential sector. It is important to note that each residential agent performs a model predictive control to perform actions such as preheating the house to avoid high-price regions, respecting comfort needs, and set-point temperature changes.

Table 2
Specifications of the computer used for the simulation process.

Component	Description
Processor	Intel Xeon W-2245 3.90 GHz
Memory	128 GB - DDR4
Hard drive size	4TB SSD
OS	Ubuntu-22.04

Table 2 provides the hardware specification of the computer used for simulation purposes. It is important to mention that most of the computational burden was linked to the residential models. These computations accounted for more than 98% of the system computation time. This problem can be alleviated through the implementation of distributed computation strategies that better represent the actual behavior of these architectures.

MARL for optimizing DRA dynamic pricing strategy. The MARL environment is developed using the OpenAI Gym API. The 11 developed residential agents are distributed between three DRAs in this environment. One DRA with three customers and the other two with four. The price limits at the aggregator level are $\pi_0 = 15$ ¢/kWh and π_{min} can be established by the DRAs within the interval $[5, 15]$ in ¢/kWh. These values will be used to build the price generator function. At the DSO level, the reward function (4) will utilize the parameters $c_1 = 20$ and $c_2 = 1.42$. These parameters come from the PAR-based form of the function proposed by [46]. Finally, as it is important to balance the terms of each DRA's reward function (9) and it is not an easy task to determine the grid cost reduction for a peak shaving achieved, $\lambda^{max} = 1$ representing the 100% of a given reward, and $\omega_1 = 1$ as well. On the other hand, for each DRA n, $\omega_2 = \sum_{k=1}^{N} \pi^n_k(a^{n,0}) y^{n,0}_k$ to normalize the second term of the rewards function with respect to the initial DRA revenue with the constant price π_0. These values are fixed for all iterations in this case study.

The proposed MARL approach starts with a learning process during 1000 episodes. Each episode comprises a coordination loop that stops after a maximum of 10 iterations between each aggregator and its customers or when threshold defined in (15) is less than 1%. Fig. 5 provides the IPPO algorithm's performance during training, presenting the aggregate reward of the different RL agents. The DRAs initially select poor actions for the parameter setting of the price function through the coordination loop. By exploiting the experience they gradually gain, the DRAs finally start improving their decision-making process, achieving higher rewards and cooperating better to decrease the PAR for the system aggregated load profile. After 500 episodes, the algorithm converges, and the system is ready for validation. More specifically, Fig. 6 evidence how each aggregator maximizes its own reward function during the training process by individually improving their decision-making rules. This Figure evidences how each agent realizes that

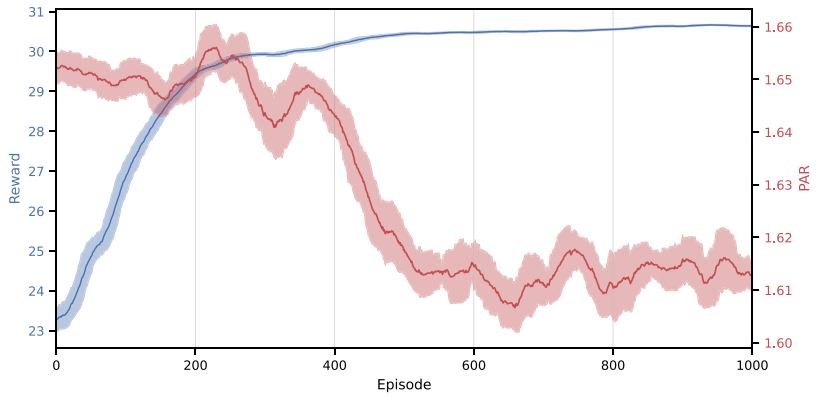


Fig. 5. Analysis of the IPPO mechanism performance during the training process.

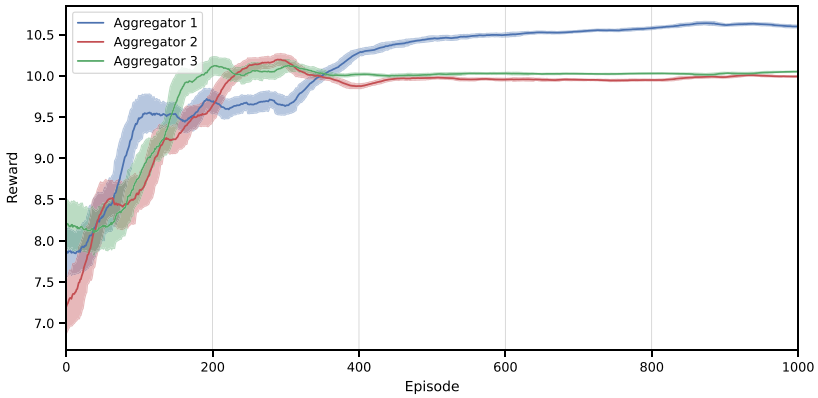


Fig. 6. Analysis of the individuals DRAs training process.

offering price discounts to the end-users allows the capitalization of the DSO's reward.

Fig. 7 shows the performance in peak reduction of the dynamic pricing mechanism for the proposed multi-aggregator system. These results demonstrate that implementing a MARL mechanism combined with SV-based reward-sharing mechanism calculation can significantly reduce peak load in a cooperative scenario. In fact, it is also possible to verify the achievement of PAR reduction, reducing the system aggregated profile's PAR from 1.9 to 1.61. Fig. 8 provides an insight into the role of each DRA in achieving the peak-shaving presented of the system aggregated consumption profile. The figure demonstrates how the coordination loop can reduce the peaks utilizing dynamic price profiles when the DRA determines the optimized parameters for the price generator function for each iteration.

Shapley-Value-based reward-sharing mechanism. Finally, to analyze the importance of combining the IPPO algorithm with the SV-based reward-sharing mechanism, in Fig. 9, a performance comparison is presented. A comparative study is conducted by implementing the same IPPO technique without utilizing the SV calculation, i.e., dividing the DSO's reward evenly between the three DRAs. In this, it is possible to verify that the fair reward-sharing mechanism improves the convergence performance of the MARL technique in terms of the convergence time, which is reduced by 29%, representing 290 episodes less for training. Calculating the marginal contributions for each DRA provides the RL agents with a better understanding of the impact of their actions on the system. This extra information helps deal with the non-stationarity problem of MARL techniques, resulting in a faster and more optimized solution.

Performance comparison. The proposed MARL-based mechanism is finally compared with a proximal decomposition approach proposed

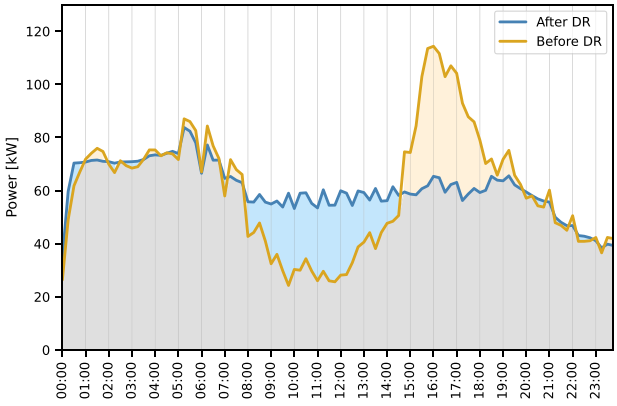
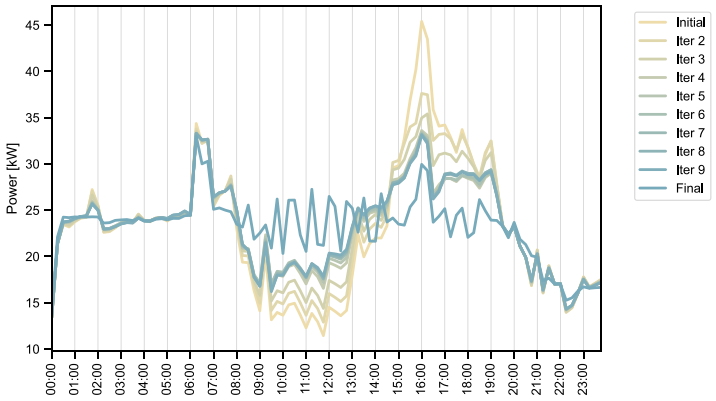
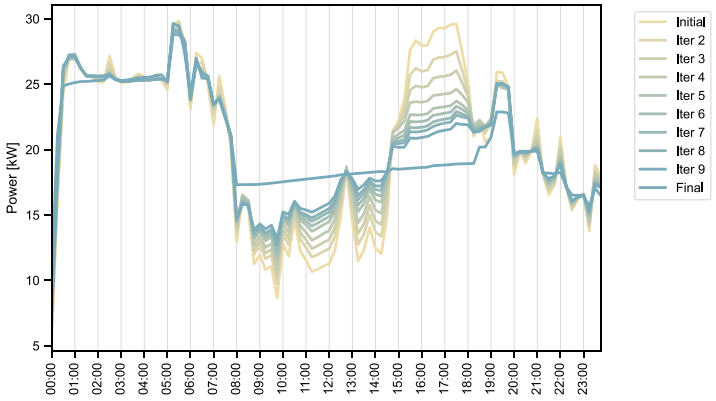


Fig. 7. Peak reduction after learning process.

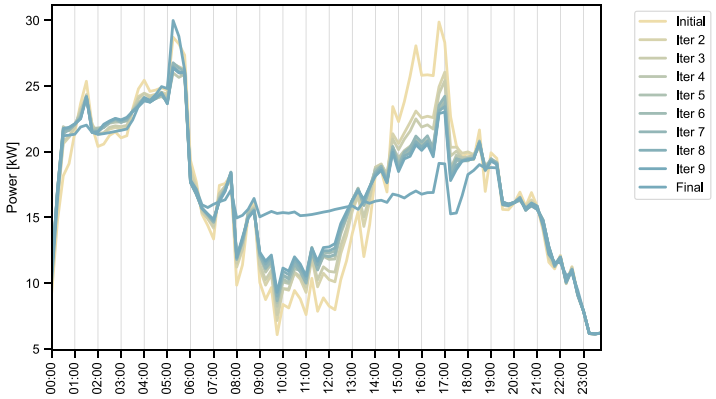
by the authors in [60]. This mechanism is applied by each DRA applying a billing mechanism proportional to the consumption plan during ten iterations. Furthermore, this mechanism is adapted to respect the price limits established in the proposed scenario for a more fair comparison. Table 3 provides the obtained results. This information demonstrates that the proximal decomposition approach can provide a higher aggregators' income from selling energy. However, the proposed MARL-based mechanism provides better results regarding PAR reduction, representing a DSO's reward 50% higher than the reward obtained with the proximal decomposition approach. This highlights the ability of the proposed model to make different aggregators cooperate in order to achieve an overall system objective.



(a) Coordination loop for DRA agent 1 with four houses.



(b) Coordination loop DRA agent 2 with four houses.



(c) Coordination loop DRA agent 3 with three houses.

Fig. 8. DRAs' coordination loops after training.

Table 3
Performance comparison between the proposed MARL-based mechanism and a proximal decomposition approach.

		DRA 1	DRA 2	DRA 3
IPPO	Income	74.6$	62.2$	55.05
	PAR	1.42	1.44	1.75
	DSO's reward	73.1%		
Proximal decomposition	Income	91.6$	80.9$	66.1$
	PAR	1.91	1.87	1.67
	DSO's reward	19.8%		

4. Conclusions

This paper proposes a cooperative price-based demand response mechanism for a multi-aggregator system, utilizing multi-agent reinforcement learning (MARL) and a Shapley Value-based reward-sharing approach. In this regard, an IPPO-based MARL architecture is employed to coordinate a set of demand response aggregator (DRA) agents, aiming to harness the flexibility potential of residential customers. The DRAs establish dynamic pricing discounts in an iterative process, where DRAs communicate their price profiles and customers adjust their consumption plans accordingly. In this win–win approach, the residential users leverage the flexibility of their controllable loads to reduce their bills, while the DRAs capitalize on this flexibility to decrease the system's aggregated peak demand. As a result, DRAs gain access to rewards

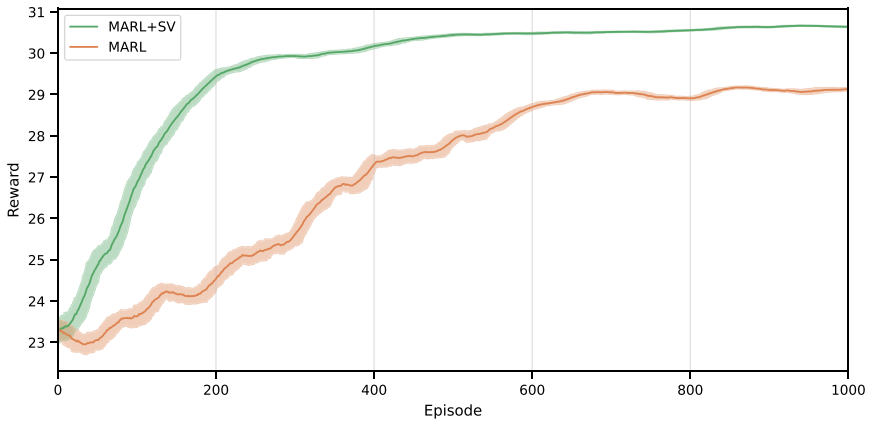


Fig. 9. MARL performance with and without the SV-based reward-sharing mechanism.

offered by the DSO for improving the peak-to-average ratio (PAR) of the daily profile. The results presented demonstrate a significant reduction in the PAR of total power demand, from 1.9 to 1.61. Furthermore, the importance of implementing the SV-based reward-sharing mechanism is shown, improving the optimization of the solution and reducing the convergence time by 29% while increasing the reward obtained by the agents by more than 5%. The proposed approach has also been compared with a mechanism based on proximal decomposition. This strategy can lead to higher income for aggregators from energy sales. However, the proposed MARL-based mechanism yields superior results in terms of PAR reduction, resulting in a DSO reward that is 50% higher than that achieved with the proximal decomposition approach.

This study has the limitation of not considering the willingness to participate from the residential users. The development of such a model would make it possible to develop strategies for adapting DRA agents according to the users who wish to participate in the program. It would also allow the evaluation of the performance of competitive scenarios among DRA agents seeking to increase the number of enrolled clients. In future studies, the development of the willingness model will be performed, as well as the consideration of a progressive change in user elasticity. Furthermore, the proposed approach will be analyzed in terms of future application by analyzing the performance of strategies to pre-train the MARL mechanism in a historical day and then evaluate the algorithm in out-of-sample days. In addition, the consideration of users' deviations from consumption plans will be explored.

CRediT authorship contribution statement

Alejandro Fraija: Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Nilson Henao:** Writing – review & editing, Validation, Methodology, Formal analysis, Conceptualization. **Kodjo Agbossou:** Supervision, Formal analysis, Conceptualization. **Sousso Kelouwani:** Writing – review & editing, Validation, Formal analysis, Conceptualization. **Michaël Fournier:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the Laboratoire des technologies de l'énergie (LTE) d'Hydro-Québec, the Natural Science and Engineering Research Council of Canada and the Foundation of UQTR.

Data availability

The data that has been used is confidential.

References

[1] S. Martinez, M. Vellei, J. Le Dréau, Demand-side flexibility in a residential district: What are the main sources of uncertainty? Energy Build. 255 (2022) 111595.

[2] S. Althaher, P. Mancarella, J. Mutale, Automated demand response from home energy management system under dynamic pricing and power and comfort constraints, IEEE Trans. Smart Grid 6 (4) (2015) 1874–1883, http://dx.doi.org/10.1109/TSG.2014.2388357.

[3] A.R. Jordehi, Optimisation of demand response in electric power systems, a review, Renew. Sustain. Energy Rev. 103 (2019) 308–319.

[4] D.A. Khan, A. Arshad, M. Lehtonen, K. Mahmoud, Combined DR pricing and voltage control using reinforcement learning based multi-agents and load forecasting, IEEE Access 10 (2022) 130839–130849.

[5] S. Burger, J.P. Chaves-Ávila, C. Batlle, I.J. Pérez-Arriaga, A review of the value of aggregators in electricity systems, Renew. Sustain. Energy Rev. 77 (2017) 395–405, http://dx.doi.org/10.1016/j.rser.2017.04.014, URL https://www.sciencedirect.com/science/article/pii/S1364032117305191.

[6] J. Stede, K. Arnold, C. Dufter, G. Holtz, S. von Roon, J.C. Richstein, The role of aggregators in facilitating industrial demand response: Evidence from Germany, Energy Policy 147 (2020) 111893, http://dx.doi.org/10.1016/j.enpol.2020.111893, URL https://www.sciencedirect.com/science/article/pii/S030142152030608X.

[7] V. Rigoni, D. Flynn, A. Keane, Coordinating demand response aggregation with LV network operational constraints, IEEE Trans. Power Syst. 36 (2) (2021) 979–990, http://dx.doi.org/10.1109/TPWRS.2020.3014144.

[8] M.A. Khan, A.M. Saleh, M. Waseem, I.A. Sajjad, Artificial intelligence enabled demand response: Prospects and challenges in smart grid environment, IEEE Access 11 (2023) 1477–1505, http://dx.doi.org/10.1109/ACCESS.2022.3231444.

[9] X. Zhang, D. Biagioni, P. Graf, J. King, Cooperative load scheduling for multiple aggregators using hierarchical admm, in: 2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference, ISGT, IEEE, 2020, pp. 1–5.

[10] R. Bagherpour, N. Mozayani, B. Badnava, Optimizing dynamic pricing demand response algorithm using reinforcement learning in smart grid, in: 2020 25th International Computer Conference, Computer Society of Iran, CSICC, 2020, pp. 1–5, http://dx.doi.org/10.1109/CSICC49403.2020.9050115.

[11] M.S. Bakare, A. Abdulkarim, M. Zeeshan, A.N. Shuaibu, A comprehensive overview on demand side energy management towards smart grids: challenges, solutions, and future direction, Energy Inform. 6 (1) (2023) 4.

[12] H. Yu, J. Zhang, J. Ma, C. Chen, G. Geng, Q. Jiang, Privacy-preserving demand response of aggregated residential load, Appl. Energy 339 (2023) 121018, http://dx.doi.org/10.1016/j.apenergy.2023.121018, URL https://www.sciencedirect.com/science/article/pii/S0306261923003823.

[13] E.J. Salazar, M. Jurado, M.E. Samper, Reinforcement learning-based pricing and incentive strategy for demand response in smart grids, Energies 16 (3) (2023) http://dx.doi.org/10.3390/en16031466, URL https://www.mdpi.com/1996-1073/16/3/1466.

[14] G. O'Brien, A. El Gamal, R. Rajagopal, Shapley value estimation for compensation of participants in demand response programs, IEEE Trans. Smart Grid 6 (6) (2015) 2837–2844, http://dx.doi.org/10.1109/TSG.2015.2402194.

[15] L.S. Shapley, Notes on the N-Person Game — II: The Value of an N-Person Game, RAND Corporation, Santa Monica, CA, 1951, http://dx.doi.org/10.7249/RM0670.

[16] S. Han, H. Wang, S. Su, Y. Shi, F. Miao, Stable and efficient Shapley value-based reward reallocation for multi-agent reinforcement learning of autonomous vehicles, in: 2022 International Conference on Robotics and Automation, ICRA, IEEE, 2022, pp. 8765–8771.

[17] R. Lowe, Y.I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, I. Mordatch, Multi-agent actor-critic for mixed cooperative-competitive environments, Adv. Neural Inf. Process. Syst. 30 (2017).

[18] M.I. Ohannessian, M. Roozbehani, D. Materassi, M.A. Dahleh, Dynamic estimation of the price-response of deadline-constrained electric loads under threshold policies, in: 2014 American Control Conference, 2014, pp. 2798–2803, http://dx.doi.org/10.1109/ACC.2014.6859473.

[19] H. Taherian, M.R. Aghaebrahimi, L. Baringo, S.R. Goldani, Optimal dynamic pricing for an electricity retailer in the price-responsive environment of smart grid, Int. J. Electr. Power Energy Syst. 130 (2021) 107004, http://dx.doi.org/10.1016/j.ijepes.2021.107004, URL https://www.sciencedirect.com/science/article/pii/S0142061521002441.

[20] S. Nojavan, K. Zare, B. Mohammadi-Ivatloo, Optimal stochastic energy management of retailer based on selling price determination under smart grid environment in the presence of demand response program, Appl. Energy 187 (2017) 449–464, http://dx.doi.org/10.1016/j.apenergy.2016.11.024, URL https://www.sciencedirect.com/science/article/pii/S0306261916316099.

[21] D. Zhang, H. Zhu, H. Zhang, H.H. Goh, H. Liu, T. Wu, Multi-objective optimization for smart integrated energy system considering demand responses and dynamic prices, IEEE Trans. Smart Grid 13 (2) (2022) 1100–1112, http://dx.doi.org/10.1109/TSG.2021.3128547.

[22] S. Datchanamoorthy, S. Kumar, Y. Ozturk, G. Lee, Optimal time-of-use pricing for residential load control, in: 2011 IEEE International Conference on Smart Grid Communications, SmartGridComm, 2011, pp. 375–380, http://dx.doi.org/10.1109/SmartGridComm.2011.6102350.

[23] L. Jia, L. Tong, Dynamic pricing and distributed energy management for demand response, IEEE Trans. Smart Grid 7 (2) (2016) 1128–1136, http://dx.doi.org/10.1109/TSG.2016.2515641.

[24] C. Feng, Z. Li, M. Shahidehpour, F. Wen, Q. Li, Stackelberg game based transactive pricing for optimal demand response in power distribution systems, Int. J. Electr. Power Energy Syst. 118 (2020) 105764, http://dx.doi.org/10.1016/j.ijepes.2019.105764, URL https://www.sciencedirect.com/science/article/pii/S0142061519326407.

[25] L.D. Collins, R.H. Middleton, Distributed demand peak reduction with non-cooperative players and minimal communication, IEEE Trans. Smart Grid 10 (1) (2019) 153–162, http://dx.doi.org/10.1109/TSG.2017.2734113.

[26] C. Silva, P. Faria, Z. Vale, Finding the trustworthy consumers for demand response events by dealing with uncertainty, in: 2021 IEEE International Conference on Environment and Electrical Engineering and 2021 IEEE Industrial and Commercial Power Systems Europe, EEEIC / I&CPS Europe, 2021, pp. 1–6, http://dx.doi.org/10.1109/EEEIC/ICPSEurope51590.2021.9584667.

[27] J.R. Vázquez-Canteli, Z. Nagy, Reinforcement learning for demand response: A review of algorithms and modeling techniques, Appl. Energy 235 (2019) 1072–1089, http://dx.doi.org/10.1016/j.apenergy.2018.11.002, URL https://www.sciencedirect.com/science/article/pii/S0306261918317082.

[28] R. Lu, S.H. Hong, X. Zhang, A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach, Appl. Energy 220 (2018) 220–230, http://dx.doi.org/10.1016/j.apenergy.2018.03.072, URL https://www.sciencedirect.com/science/article/pii/S0306261918304112.

[29] S. Zhong, X. Wang, J. Zhao, W. Li, H. Li, Y. Wang, S. Deng, J. Zhu, Deep reinforcement learning framework for dynamic pricing demand response of regenerative electric heating, Appl. Energy 288 (2021) 116623, http://dx.doi.org/10.1016/j.apenergy.2021.116623, URL https://www.sciencedirect.com/science/article/pii/S0306261921001586.

[30] A. Ghasemkhani, L. Yang, Reinforcement learning based pricing for demand response, in: 2018 IEEE International Conference on Communications Workshops, ICC Workshops, 2018, pp. 1–6, http://dx.doi.org/10.1109/ICCW.2018.8403783.

[31] B.-G. Kim, Y. Zhang, M. van der Schaar, J.-W. Lee, Dynamic pricing for smart grid with reinforcement learning, in: 2014 IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS, 2014, pp. 640–645, http://dx.doi.org/10.1109/INFCOMW.2014.6849306.

[32] Y. Du, F. Li, Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning, IEEE Trans. Smart Grid 11 (2) (2020) 1066–1076, http://dx.doi.org/10.1109/TSG.2019.2930299.

[33] R. Li, A.J. Satchwell, D. Finn, T.H. Christensen, M. Kummert, J. Le Dréau, R.A. Lopes, H. Madsen, J. Salom, G. Henze, K. Wittchen, Ten questions concerning energy flexibility in buildings, Build. Environ. 223 (2022) 109461, http://dx.doi.org/10.1016/j.buildenv.2022.109461, URL https://www.sciencedirect.com/science/article/pii/S0360132322006928.

[34] K.T. Ponds, A. Arefi, A. Sayigh, G. Ledwich, Aggregator of demand response for renewable integration and customer engagement: Strengths, weaknesses, opportunities, and threats, Energies 11 (9) (2018) http://dx.doi.org/10.3390/en11092391, URL https://www.mdpi.com/1996-1073/11/9/2391.

[35] S. Zheng, Y. Sun, B. Li, B. Qi, K. Shi, Y. Li, Y. Du, Bargaining-based cooperative game among multi-aggregators with overlapping consumers in incentive-based demand response, IET Gener. Transm. Distrib. 14 (6) (2020) 1077–1090.

[36] L. Canese, G.C. Cardarilli, L. Di Nunzio, R. Fazzolari, D. Giardino, M. Re, S. Spanò, Multi-agent reinforcement learning: A review of challenges and applications, Appl. Sci. 11 (11) (2021) 4948.

[37] J. Wang, W. Xu, Y. Gu, W. Song, T.C. Green, Multi-agent reinforcement learning for active voltage control on power distribution networks, Adv. Neural Inf. Process. Syst. 34 (2021) 3271–3284.

[38] M. Roesch, C. Linder, R. Zimmermann, A. Rudolf, A. Hohmann, G. Reinhart, Smart grid for industry using multi-agent reinforcement learning, Appl. Sci. 10 (19) (2020) 6900.

[39] H. Kazmi, J. Suykens, A. Balint, J. Driesen, Multi-agent reinforcement learning for modeling and control of thermostatically controlled loads, Appl. Energy 238 (2019) 1022–1035, http://dx.doi.org/10.1016/j.apenergy.2019.01.140, URL https://www.sciencedirect.com/science/article/pii/S0306261919301564.

[40] J. van Tilburg, L.C. Siebert, J.L. Cremer, MARL-iDR: Multi-agent reinforcement learning for incentive-based residential demand response, 2023, arXiv preprint arXiv:2304.04086.

[41] L. Hou, Y. Li, J. Yan, C. Wang, L. Wang, B. Wang, Multi-agent reinforcement mechanism design for dynamic pricing-based demand response in charging network, Int. J. Electr. Power Energy Syst. 147 (2023) 108843, http://dx.doi.org/10.1016/j.ijepes.2022.108843, URL https://www.sciencedirect.com/science/article/pii/S0142061522008390.

[42] A. Shojaeighadikolaei, A. Ghasemi, K.R. Jones, A.G. Bardas, M. Hashemi, R. Ahmadi, Demand responsive dynamic pricing framework for prosumer dominated microgrids using multiagent reinforcement learning, in: 2020 52nd North American Power Symposium, NAPS, IEEE, 2021, pp. 1–6.

[43] E.J. Salazar, V. Rosero, J. Gabrielski, M.E. Samper, Demand response model: A cooperative-competitive multi-agent reinforcement learning approach, Eng. Appl. Artif. Intell. 133 (2024) 108273, http://dx.doi.org/10.1016/j.engappai.2024.108273, URL https://www.sciencedirect.com/science/article/pii/S0952197624004317.

[44] J. Wang, Q. Huang, W. Hu, J. Li, Z. Zhang, D. Cai, X. Zhang, N. Liu, Ensuring profitability of retailers via Shapley value based demand response, Int. J. Electr. Power Energy Syst. 108 (2019) 72–85.

[45] F. Khavari, A. Badri, A. Zangeneh, Energy management in multi-microgrids via an aggregator to override point of common coupling congestion, IET Gener. Transm. Distrib. 13 (5) (2019) 634–642.

[46] F. Etedadi, S. Kelouwani, K. Agbossou, N. Henao, F. Laurencelle, Consensus and sharing based distributed coordination of home energy management systems with demand response enabled baseboard heaters, Appl. Energy 336 (2023) 120833, http://dx.doi.org/10.1016/j.apenergy.2023.120833, URL https://www.sciencedirect.com/science/article/pii/S0306261923001976.

[47] J. Li, Interdependent relationships in game theory: A generalized model, 2016, arXiv preprint arXiv:1601.00176.

[48] C.L. Dewangan, S. Singh, S. Chakrabarti, K. Singh, Peak-to-average ratio incentive scheme to tackle the peak-rebound challenge in TOU pricing, Electr. Power Syst. Res. 210 (2022) 108048, http://dx.doi.org/10.1016/j.epsr.2022.108048, URL https://www.sciencedirect.com/science/article/pii/S0378779622002735.

[49] A. Fraija, N. Henao, K. Agbossou, S. Kelouwani, M. Fournier, S.H. Nagarsheth, Deep reinforcement learning based dynamic pricing for demand response considering market and supply constraints, Smart Energy (2024) 100139, http://dx.doi.org/10.1016/j.segy.2024.100139, URL https://www.sciencedirect.com/science/article/pii/S2666955224000091.

[50] I. Jendoubi, F. Bouffard, Multi-agent hierarchical reinforcement learning for energy management, Appl. Energy 332 (2023) 120500.

[51] P. Atrazhev, P. Musilek, It's all about reward: Contrasting joint rewards and individual reward in centralized learning decentralized execution algorithms, Systems 11 (4) (2023) 180.

[52] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, 2017, arXiv preprint arXiv:1707.06347.

[53] K. Su, Z. Lu, Decentralized policy optimization, 2022, arXiv preprint arXiv:2211.03032.

[54] D. Azuatalam, W.-L. Lee, F. de Nijs, A. Liebman, Reinforcement learning for whole-building HVAC control and demand response, Energy AI 2 (2020) 100020, http://dx.doi.org/10.1016/j.egyai.2020.100020, URL https://www.sciencedirect.com/science/article/pii/S2666546820300203.

[55] D. Toquica, K. Agbossou, N. Henao, R. Malhamé, S. Kelouwani, F. Amara, Prevision and planning for residential agents in a transactive energy environment, Smart Energy 2 (2021) 100019.

[56] J.A. Dominguez, K. Agbossou, N. Henao, S.H. Nagarsheth, J. Campillo, L. Rueda, Distributed stochastic energy coordination for residential prosumers: Framework and implementation, Sustain. Energy Grids Netw 38 (2024) 101324, http://dx.doi.org/10.1016/j.segan.2024.101324, URL https://www.sciencedirect.com/science/article/pii/S2352467724000535.

[57] R. Deng, Z. Yang, J. Chen, M.-Y. Chow, Load scheduling with price uncertainty and temporally-coupled constraints in smart grids, IEEE Trans. Power Syst. 29 (6) (2014) 2823–2834, http://dx.doi.org/10.1109/TPWRS.2014.2311127.

[58] G. Scutari, D.P. Palomar, F. Facchinei, J.-S. Pang, Monotone games for cognitive radio systems, Distributed Decis. Mak. Control. (2012) 83–112.
[59] A. Fraija, K. Agbossou, N. Henao, S. Kelouwani, M. Fournier, S.S. Hosseini, A discount-based time-of-use electricity pricing strategy for demand response with minimum information using reinforcement learning, IEEE Access 10 (2022) 54018–54028, http://dx.doi.org/10.1109/ACCESS.2022.3175839.
[60] H.K. Nguyen, J.B. Song, Z. Han, Distributed demand side management with energy storage in smart grid, IEEE Trans. Parallel Distrib. Syst. 26 (12) (2015) 3346–3357, http://dx.doi.org/10.1109/TPDS.2014.2372781.
[61] N. Henao, M. Fournier, S. Kelouwani, Characterizing smart thermostats operation in residential zoned heating systems and its impact on energy saving metrics, in: Proceedings of ESim 2018, the 10th Conference of IBPSA-Canada, 2018, pp. 17–25.