UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

GÉNÉRATION DE POLITIQUES TRANSACTIONNELLES POUR LES AGRÉGATEURS DE RÉPONSE À LA DEMANDE

THÈSE PRÉSENTÉE
COMME EXIGENCE PARTIELLE DU
DOCTORAT EN GÉNIE ÉLECTRIQUE

PAR
ALEJANDRO JOSÉ FRAIJA OCHOA

avril, 2024

UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

DOCTORAT EN GÉNIE ÉLECTRIQUE (Ph.D.)

Direction de recherche :

Prof. Kodjo Agbossou — Directeur de recherche
Université du Québec à Trois-Rivières

Prof. Sousso Kelouwani — Codirecteur de recherche
Université du Québec à Trois-Rivières

Jury d'évaluation

François Nougarou — Président du jury
Université du Québec à Trois-Rivières

Ricardo Izquierdo — Évaluateur externe
École de technologie supérieure | Université du Québec

César Duarte — Évaluateur externe
Universidad Industrial de Santander

Prof. Kodjo Agbossou — Directeur de recherche
Université du Québec à Trois-Rivières

Prof. Sousso Kelouwani — Codirecteur de recherche
Université du Québec à Trois-Rivières

Thèse soutenue le 6 mars 2024

Abstract

The increase in energy needs of the growing global population and the concern with its associated Greenhouse Gas emissions cause significant challenges for conventional power systems. In this regard, the smart grid concept is proposed as a key enabler for clean energy generation and efficient energy consumption. Under the smart grid paradigm, the emergence of transactive energy systems brings about a remarkable opportunity for realizing a modernized power grid through enhanced Energy Management Systems. Particularly, these new systems offer innovative Demand Response (DR) programs in order to improve energy efficiency and flexibility and facilitate renewable resources and energy storage integration. This is achieved by leveraging advanced metering infrastructure, two-way communication networks, and distributed control systems. The smart grid also frames a group of mechanisms for DR characterized by generating incentives or pricing policies in an adaptive and more real-time manner. Consumers can react by modifying their load profiles in order to minimize energy costs while maintaining comfort desires. On the grid side, the operator can manage system congestion and minimize operational costs by reducing the peak demand and deferring the construction of new power plants and power delivery systems. However, these DR programs face significant challenges in terms of modeling and management of decision-support information in dynamic and non-homogeneous environments. Indeed, the incomplete information on the dynamics of the behind-the-meter resources, the inherent issues of user data confidentiality, the potential failures in the communication channels, and the emergence of intelligent loads (including storage) create a complex and uncertain environment for the decision-making process.

As a result, a new entity is emerging, the demand response aggregator. This aggregator acts as a mediator between consumers and the electricity market to explore the flexibility opportunities offered by the residential sector. This new entity will then seek to offer benefits to both parties (distributors and users) by exploiting the policies of demand response programs. The mentioned role translates into an interaction between players seeking to maximize their gains and thus ends up being framed by game theory. However, the various sources of uncertainty mentioned above considerably complicate the process of generating optimal policies. With this in mind, reinforcement learning methods emerge, offering the possibility of managing uncertainty through a trial-and-error process. In other words, this approach takes advantage of the interactions between the various players in the system, in order to achieve an optimized generation of transactive policies.

This thesis proposes to develop an automated agent (meeting the needs of network managers) for generating optimized transactive policies through interactions in a residential environment. The proposed approach considers a transactive environment composed of rational residential agents and a demand response aggregator agent interacting in a game theoretic framework. The aggregator's adaptability and ability to handle uncertainty are considered through reinforcement learning techniques. The results demonstrate the effectiveness of the proposed method in managing residential consumption. The aggregator agent is able to offer economic incentives to users through the development of pricing policies while respecting users' privacy, in order to exploit the potential of residential flexibility.

Résumé

L'augmentation des besoins en énergie de la population mondiale et les préoccupations liées aux émissions de gaz à effet de serre qui y sont associées posent des défis importants aux systèmes électriques conventionnels. Dans ce contexte, le concept de réseau intelligent (SG) est proposé comme un outil clé pour la production d'énergie propre et la consommation d'énergie efficace. Dans le cadre du paradigme du SG, l'émergence de systèmes énergétiques transactionnels offre une opportunité remarquable de moderniser le réseau électrique grâce à l'évolution des systèmes de gestion de l'énergie. En particulier, ces systèmes offrent des programmes innovants de réponse à la demande (DR) afin d'améliorer l'efficacité et la flexibilité énergétiques et de faciliter l'intégration des ressources renouvelables et du stockage. Pour ce faire, ils s'appuient sur une infrastructure de comptage avancée, des réseaux de communication bidirectionnels et des systèmes de contrôle distribués. Le SG encadre également un groupe de mécanismes pour le DR caractérisés par la génération d'incitatifs ou de politiques de tarification d'une manière adaptative et en temps réel. Les consommateurs peuvent réagir en modifiant leurs profils de charge afin de minimiser les coûts énergétiques tout en maintenant leurs désirs de confort. Du côté du réseau, l'opérateur peut gérer les contraintes opérationnelles du réseau et minimiser les coûts opérationnels en réduisant la demande de pointe et en reportant la construction de nouvelles centrales électriques et de systèmes de distribution d'énergie. Cependant, ces programmes de réduction de la consommation sont confrontés à des défis importants en termes de modélisation et de gestion des informations dans des environnements dynamiques et non homogènes. En effet, les informations incomplètes sur la dynamique des ressources derrière le compteur, les problèmes inhérents à la confidentialité des données des utilisateurs, les défaillances potentielles des canaux de communication et l'émergence de charges intelligentes (y compris le stockage) créent un environnement complexe et incertain pour le processus de prise de décision.

En conséquence, une nouvelle entité émerge, l'agrégateur de réponse à la demande (DRA). Cet entité joue le rôle de médiateur entre les consommateurs et le marché de l'électricité afin d'explorer les possibilités de flexibilité offertes par le secteur résidentiel. Cette nouvelle entité cherchera alors à offrir des avantages aux deux parties (distributeurs et utilisateurs) en exploitant les politiques des programmes de DR. Le rôle mentionné se traduit par une interaction entre des acteurs cherchant à maximiser leurs gains et finit donc par être encadré par la théorie des jeux. Cependant, les différentes sources d'incertitude mentionnées ci-dessus compliquent considérablement le processus de génération de politiques optimales. Dans cette optique, les méthodes d'apprentissage par renforcement (RL) apparaissent, offrant la possibilité de gérer l'incertitude par le biais d'un processus itératif d'action-récompense. En d'autres termes, cette approche tire parti des interactions entre les différents acteurs du système, afin de parvenir à une génération optimisée de politiques transactionnelles.

Cette thèse propose de développer un agent DRA automatisé pour générer des politiques transactionnelles optimisées par le biais d'interactions dans un environnement résidentiel. L'approche proposée considère un environnement transactionnel composé d'agents résidentiels rationnels et l'agent DRA interagissant dans un schéma de théorie des jeux. L'adaptabilité de l'agrégateur et sa capacité à gérer l'incertitude sont prises en compte grâce à des techniques de RL. Les résultats démontrent l'efficacité des approches proposées dans la gestion de la consommation résidentielle. L'agent DRA est capable d'offrir des incitations économiques aux utilisateurs à travers le développement de politiques de tarification tout en respectant la vie privée des utilisateurs, afin d'exploiter le potentiel de flexibilité résidentielle.

Acknowledgment

I would like to sincerely thank my research advisor, Prof. Kodjo Agbossou, a member of the Department of Electrical and Computer Engineering at UQTR and Director of the Laboratoire d'Innovation et de Recherche en Énergie Intelligente (LIREI), for entrusting me with this project, for the quality of his supervision, and for his advice, support, and encouragement. I would also like to thank my co-advisor, Prof. Sousso Kélouwani, a member of the Mechanical Engineering Department at UQTR, for his help, his scientific support, his invaluable contributions, and his availability throughout this thesis.

I would like to express my special gratitude to DR. Nilson Henao, research agent at LIREI, for his technical support, his participation in the development of this research project, and for always pushing me to improve the quality of my work. I would also like to thank Michael Fournier, a researcher at Hydro-Québec's Laboratoire de Technologies de l'Énergie (LTE), for his scientific participation, and for the time he devoted to our exchanges, meetings, and revisions of documents relating to the research work.

This work would not have been possible without the help of various institutes and organizations who, through their financial or technical support, recognized my work and placed their trust in me: Hydro-Québec's LTE, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Institut de Recherche sur l'Hydrogène (IRH) and the Fondation de l'UQTR.

Deseo dedicarle esta tesis a:

Mi madre,

Por el amor, los consejos y el apoyo que en vida me diste. Por haber sido la inspiración y la luz que me guió en mi camino. Porque fuiste tú la que me enseño que lo mas valioso en la vida es el estudio. Este trabajo es la muestra de que gracias a la huella que dejaste en mi, hoy puedo decir que soy un triunfador. Sientete orgullosa porque tu hijo ahora es un doctor.

Mi esposa e hijos,

Ustedes, siendo la mayor motivación en mi vida, fueron mi fuerza y mi apoyo incondicional para poder lograr alcanzar esta gran victoria. Quiero que sepan que todo el esfuerzo y dedicacion puesto en este trabajo ha sido por y para ustedes, y que cada letra y punto de esta tesis esta puesto en sus nombres. Es a ustedes a quienes mas les tengo que agradecer. Siempre serán mi mayor logro.

Table of contents

List of figures

List of Acronyms

Notation	Description
DR	Demand Response.
DRA	Demand Response Aggregator.
DSM	Demand-side Management.
DSO	Distributed System Operator
ESH	Electric Space Heating
HEMS	Home Energy Management System.
IPPO	Independent Proximal Policy Optimization.
LF	Load Factor.
MARL	Multi-Agent Reinforcement Learning.
PAR	Peak-to-Average-Ratio.
PPO	Proximal Policy Optimization.
RL	Reinforcement learning.
SG	Smart Grid.
SV	Shapley Value.
ToU	Time-of-Use.

Chapter 1 - Introduction

Currently, different environmental challenges have raised the need to develop different strategies to overcome climate problems. In that sense, power systems have been experiencing a rapid evolution due to the modus operandi of generation-based electrical energy systems [1]. In terms of economic limitations and environmental considerations, the traditional implementation of large centralized generators within a monopoly is considered non-optimal and unsustainable. Furthermore, the environmental considerations have resulted in new governmental goals, translated into carbon taxes, emission limits, and fast implementation of renewable energy technologies [2]. For instance, according to Hydro-Québec's 2021 report, nearly 40% of the province of Québec's energy consumption is demanded by the residential sector [3]. In addition, the exposure to long winter periods makes the consumption of thermal loads represent more than 70% of residential consumption [4], as it is presented in Figure 1.1. Because of this, although the data shows that energy production is sufficient to meet the needs of grid users, it is possible that on certain days during the winter period, consumption demand exceeds production during peak consumption hours [5].

Considering the above, the idea of continuing with a one-way transaction system has become an obsolete concept. The growth of renewable electricity generation has opened the opportunity to inject energy into the grid in a decentralized way, resulting in the creation of new electricity markets that allow transactions even between users [6]. However, this generates an increasing challenge since maintaining the electric power balance while respecting the constraints of the energy grid is becoming a bigger problem. On the other hand, the integration of information and communication technologies as the focus of the Internet of Things has allowed the development of a new concept called

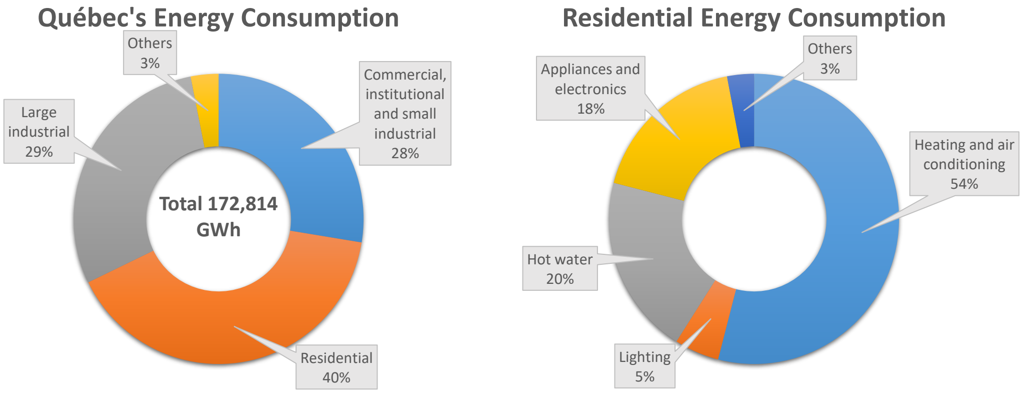


Figure 1.1 Québec's Energy consumption in 2021, [3].

smart grid (SG). This concept seeks to use these technologies to achieve great energy savings and improve the agility, reliability, efficiency, security, economy, sustainability, and environmental friendliness of the power grid [7].

An important focus in implementing the SG concept is the consideration of the role of users in energy management; this is known as Demand Response (DR). DR is a change in the electric consumption by customers from their normal consumption behavior. This is a response to changes in the electricity price or to an incentive payment designed by the operator. The idea is to induce lower electricity use at times of high wholesale market prices or when the system reliability is jeopardized [8]. The contribution of DR programs in the power systems comes from economic and reliability aspects. From the economic side, DR can shift the energy consumption from high-cost to low-cost periods, which results in a reduction of cost generation. And from the system reliability, demand response can help in the challenging task of maintaining the system frequency, and the balance between supply and demand [9].

In order to obtain an optimal behavior of the grid, a new entity called DR aggregator

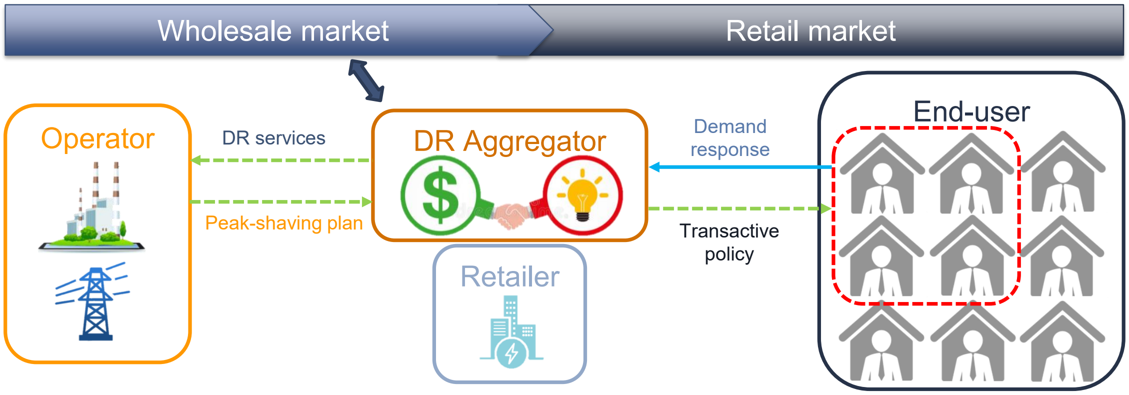


Figure 1.2 The role of a demand response aggregator, [10], [11].

(DRA) has been suggested. A DRA in the electricity market is considered a mediator or a third-party intermediate between power market participants and electricity prosumers and consumers, as presented in Figure 1.2. These aggregators offer customers contracts, enabling them to participate directly in the wholesale market while at the same time supplying services to operators that improve the reliability of the network. This is due to the fact that DRA realizes the existence of flexibility opportunities related to the users' ability to manage loads/generations [12]. They detect these opportunities by managing energy consumption that consumers carry out during critical periods of the day, such as peak hours, or the change in their behavior when they are exposed to periods with cheaper electricity. The DRA capitalizes on these opportunities to be able to enter the wholesale electricity markets. In this way, the DR can be bought by the operators and other market entities as ancillary services, capacity reserves, or balancing provisions [10].

It is possible to summarize the above by saying that the definition of this SG has allowed the development of both technical and economic instruments for the power grid. This new model enhances the participation of different actors through the emphasis on information exchange and distributed optimization [13]. These actors can be considered agents deployed in a multi-agent environment, capable of negotiating, coordinating, or cooperating according to the resources they offer or need. As a result, the agents can reach a system equilibrium by solving their individual optimization problems. This incredible interaction, called transactive energy, is possible thanks to the bidirectional communication of the new network. This allows the easy integration of renewable resources, distributed energy resources, and new technologies such as electric vehicles. Moreover, it creates a further need for the development of new innovative methodologies to overcome the new challenges linked to the new management of electricity consumption [14]. According to the framework presented, the problematic of this research work will be described below.

1.1 Problematic

Human nature is one of the major problems affecting the design of DR markets. When analyzing the behavior of large energy consumers, it can be observed that they present a rational response with respect to maximizing their profits. However, small consumers, such as residential users, do not show the same rationality. This is because the preferences of these users differ greatly and in many cases minimizing their bill may not be in their best interest [2]. Furthermore, authors in [15] conducted an study about the price elasticity of the houses. This research work found that there not exist a linear relationship between the consumption change and the price change, as it is conventionally assumed. Instead, they found that the consumption change in response to any price change will be similar for any magnitude of that price change. Because of this, although the increasing immersion of smart devices at the residential level has facilitated the process of exploiting the flexibility potential of users, the generation of optimal pricing policies remains a challenge for DR programs [16].

Therefore, the DRA is in charge of overcoming these problems to optimize the detection of flexibility opportunities presented in the network and then capitalize on them. Thus, it is necessary to develop tools for the DRA that allow us to define optimized transactive policies in order to mitigate consumption peaks from the residential sector [11]. These policies should be adapted through the characterization of the residential users that will interact with the DRA. As a result, several challenges arise that may affect the generation of these transactive policies, linked to the anticipation of the responsive behavior of residential energy consumers and grid operators' needs. It is possible to summarize them as follows:

- The uncertainties problems due to the partial observability of DR programs have resulted in excessive considerations related to access to user information for optimal policy making [17]. So, the generation of optimized transactive price policies while avoiding impacts on customer privacy is still a challenge. This approach will prevent overconfidence in the information received from users, thus avoiding opportunities for dishonest reporting and cheating the system. At the same time, it will increase users' interest in participating in the program.
- The lack of information from users would affect the convergence of the implemented methods. In addition, the characterization of consumer behavior in response to price becomes a slow process. The challenge of ensuring convergence to a near-to-optimal point while reducing the convergence time is important to guarantee the viability of future implementations of decision-making strategies in price policy generation.
- The needs of the supplier side may differ as they may be affected not only by economic aspects (the operational and power generation costs) but also by the physical constraints of the system. The vast majority of studies consider peak

shaving as a sufficient solution to improve network performance. However, these strategies rarely take into account these real needs, which can negatively affect the economic optimization of DRA decisions.

- Finally, it is necessary to consider the increase in the complexity of the system as the number of these aggregators increases. More aggregators indicate greater difficulty in scheduling the load for a utility company [18]. Moreover, aggregators that are part of different companies must cooperate to achieve global objectives while maximizing their profitability. This evidences the need for the development of cooperative models in the process of implementing DR programs.

1.2 Objectives and contributions

In response to the problems presented, this project has as an objective the proposition of strategies for generating optimized transactive policies based on intelligent agents, anticipating the individual and collective behavior of residential energy consumers. The following three specific objectives are defined as follows:

1. Generate mechanisms for the DRA to define optimized transactive policies, avoiding impacts on customer privacy and increasing user interest in participating in the DR program.
2. Formulate strategies to consider market and system constraints in the transactive policy generation process.
3. Develop a multiagent system to establish a cooperative method for a set of DRAs to achieve an overall system objective while maximizing their own profits.

The achievement of these objectives will result in the accomplishment of the following three main contributions:

1. The proposition of a method for the DRA to generate optimized transactive policies for residential customers utilizing their response to price policies and dealing with uncertainties related to the lack of domestic information.
2. The proposition of strategies for the generation of optimized transactional policies based on prices, integrating market and system constraints.
3. The proposition of a cooperative multiagent system, which will enable the management of a set of DRA to reach a global system objective while maximizing their profits.

The novelty of this study lies in the generation of pricing policies that take user privacy into account. Roughly speaking, the state of the art presents different approaches used for the generation of these policies. However, in order to optimize the transactive policy generation process, the proposed methods need specific information from consumers, which can result in two problems. The first is the loss of user interest in participating in DR programs. The second is placing excessive reliance on the information provided by customers, giving them the opportunity to cheat the system.

By choosing this approach, new challenges appear linked to the lack of information available for the generation of transactive policies. For this reason, throughout this thesis, we seek to answer different aspects, such as:

- The minimum information required to guarantee the optimization of pricing policies.
- The system conditions to ensure convergence to a near-to-optimal point or Nash equilibrium.
- The convergence time of the proposed algorithms and the proposal of methods for their reduction (if necessary).

- The implementation of fair reward mechanisms in cooperative DRA approaches.

1.3 Methodology

In order to address the discussed problem, an approach based on data-driven mechanisms is proposed. This approach exploits the interaction between the DRA agent and the residential agents for the generation of transactive policies. In this sense, a three-stage analysis is performed to achieve the objectives presented above. The first one consists of a literature search considering the proposed problem for understanding and mastering the notions linked to the domain of interest. This will be done at the same time as the residential agent models are developed so that they can interact with the DRA agents. In the second stage, the limitations and difficulties of the existing approaches will be analyzed, taking into account the requirements of the problem addressed. In the third phase, the proposal that provides an appropriate solution to the research problem is determined. Finally, the fourth phase is the performance validation of the selected mechanisms through the utilization of simulation strategies. An illustration summarizing the explained methodology followed is presented in Figure 1.3.

The exploratory study carried out allowed us to identify the methodologies used for the generation of transactive policies within the framework of DR programs. Firstly, it has shown us the need for the development of multi-agent systems to frame the interactions between the different actors of the DR program. Also, the literature review has shown the growing popularity of reinforcement learning (RL) methods for different SG applications. This is due to their ability to deal with the inherent uncertainty of DR programs, linked to the lack of information in the process of generating transactive policies. Among the advantages of this approach to solving our problem are the following:

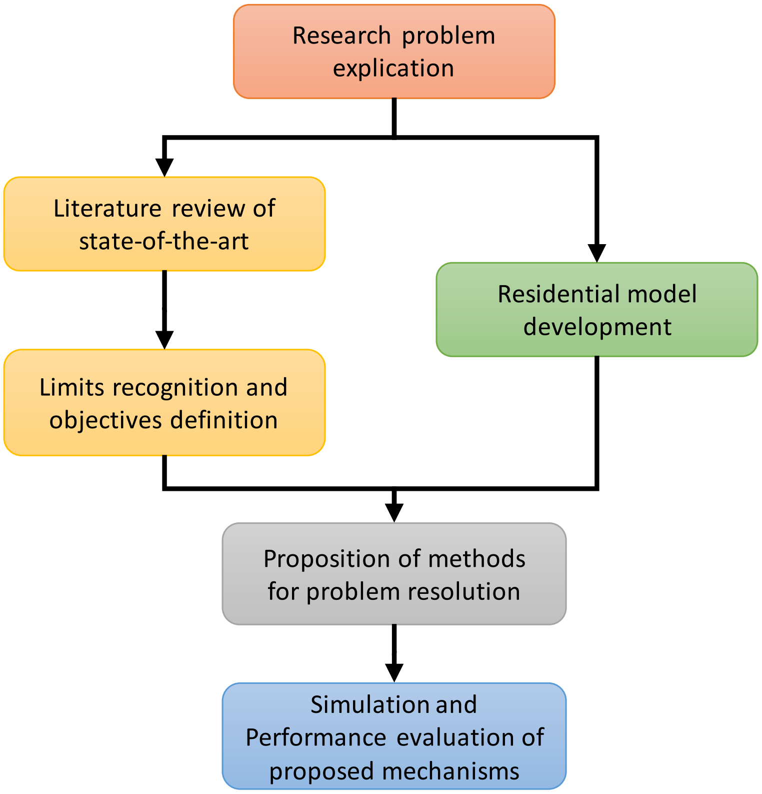

Research problem explication
Literature review of state-of-the-art
Residential model development
Limits recognition and objectives definition
Proposition of methods for problem resolution
Simulation and Performance evaluation of proposed mechanisms

Figure 1.3 The research methodology.

- This method can be used to solve very complex problems that cannot be solved with conventional techniques, as it is the case of defining an optimized price policy while ensuring customers' privacy.
- The ability to characterize the price response behavior of users through a trial and error process.
- Once the training process is done, this method can correct errors during deployment.
- It is an adaptive method that can correct its decision-making process according to changes in the environment. In this case, changes in user response patterns.
- The possibility of dealing with non-convex optimization problems.

It is clear that the implementation of RL algorithms has been explored in the SG context, especially in areas such as electric vehicles. However, certain aspects need to be analyzed for the proposed scenario to guarantee convergence to a near-to-optimal point in our multi-agent system. Therefore, the proposed approach seeks to propose a multi-agent system architecture in which a DRA agent will be able to define an optimized price policy for a given set of residential agents. This policy will aim to exploit the flexibility potential of residential agents in order to achieve an overall objective for the energy network.

Thus, to achieve the objectives of this thesis project, following the proposed three-stage approach, the following activities will be done: First, we will develop models for the controllable loads that the residential agents will be able to use as a source of flexibility. Then, a behavioral model will be built for these agents, and a control mechanism will be stated using the controllable loads to obtain a responsive behavior to stimuli (transactive policies, weather, etc.). Finally, we will analyze how the DRA can define optimized transactive policies according to the response in consumption of a given set of residential agents. The DRA will characterize the responsive behavior of the end-users by learning from their interactions. A general automated sequence for the implementation of DR programs is presented in Figure 1.4.

1.4 Assumptions

- It is assumed that the coordination mechanism is fast enough not to interfere with the execution time of the DR program. This means that the DRA will define and communicate the transactive policy in time for customers to adjust their consumption plans before the start of the next 24-hour bidding period.
- In the communication system, it is assumed that the aggregator agent is fair in sending the same information to all the residential agents. In addition, the market

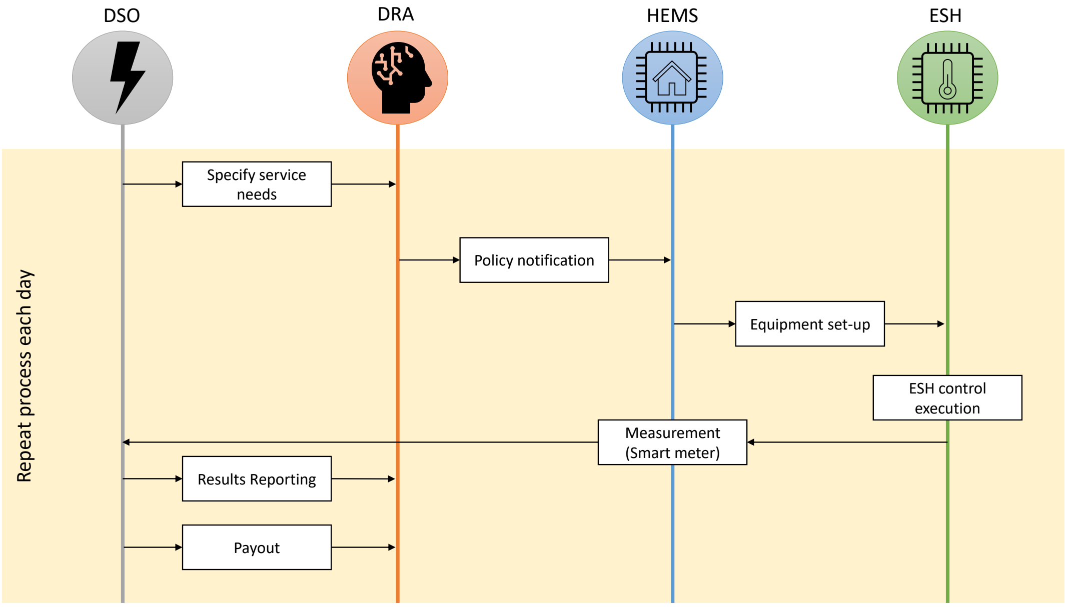


Figure 1.4 Automatic sequence for the DR program in a day-ahead market.

works in such a way that it does not allow the exchange of information between the latter, with the purpose of abusing this extra information, affecting the performance of the established DR mechanism.

- Finally, it is important to highlight the adopted assumption of consumer economic rationality, which is the basis for the implementation of DR programs concerning the price signals or incentives established. Furthermore, according to their flexibility potential, the residential agents are able to react optimally, according to the consumer's rationality.

1.5 Manuscript plan

This thesis document is composed of five chapters structured as follows:

- Chapter 1: General information about the thesis problem. In this chapter, the problem, objectives, and methodology are clearly defined.

- Chapter 2: This chapter presents the state of the art related to the generation of transactive policies. General aspects of DR programs are initially presented. The approaches used for the generation of transactive policies are discussed. Finally, a synthesis of the methods used for the generation of pricing policies in the residential context is presented.
- Chapter 3: This chapter presents the articles that have been dedicated to each objective in three main sections. Initially, it presents the development of the multi-agent architecture to ensure convergence. As a result, a first approach for the generation of pricing policies is suggested. Next, market and supply-side constraints are taken into account when proposing a DR program. Finally, a cooperative scheme between different DRAs is developed, using fair mechanisms based on the contribution of each aggregator in achieving a given global objective.
- Chapter 4: An in-depth discussion is presented to analyze the exhibited results from the previous chapter. A discussion of new opportunities and challenges that can be investigated in terms of further research subjects is emphasized for the generation of optimized price policies.
- Chapter 5: Presents conclusions and recommendations. A synthesis of the work performed is presented, ending with recommendations that will allow future improvement of the proposed methods.

Chapter 2 - State-of-the-art

2.1 Demand Side Management

Demand-side management (DSM) is a very important concept for load management that has been considered an effective tool for different tasks in the transformation of the traditional power grid into the SG. DSM provides advantages in different areas of the electricity grid, where the most popular ones are the liberalization of the electricity market, the balance in real-time of the electricity demand and supply, the improvement of control management strategies, the reduction of the energy consumption and increasing the opportunities for the implementation of decentralized energy resources, and electric vehicles [19]. The main goal of DSM is to use power-saving technologies, monetary incentives, and electricity tariffs to mitigate the energy consumption peaks, instead of increasing the generation capacity or strengthening the distribution and transmission of the grid [20].

Taking into account the above, DSM can be considered as modifications of the energy consumption pattern from the demand side to enhance the electrical energy system's efficiency and operation. These activities can be used to classify DSM techniques in load shifting, peak clipping, conservation, load building, valley filling, and flexible load shaping [21], as presented in Figure 2.1.

The literature further classifies these DSM activities into energy efficiency, strategic load growth, and DR. In the case of energy efficiency, the goal is to reduce the energy required for the provision of services or products, which can be achieved by applying load conservation techniques. On the other side, strategic load growth aims to increase the load level through electrification by applying load growth techniques. However, the modern

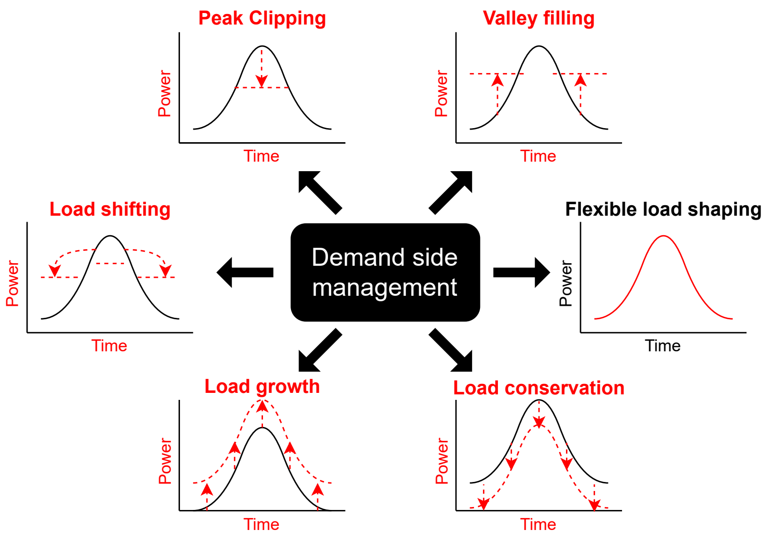


Figure 2.1 Demand side management techniques, [21].

emphasis on defining the active role of the demand side is more focused on improving the efficient and effective use of electricity, especially in resource-constrained regions. So, these techniques are not currently a target. Finally, DR activities have the objective of changing the standard consumption patterns of customers in response to changes in the price of electricity over time or to incentive payments. These changes are designed to induce lower electricity use, especially when high wholesale market prices appear or when the system reliability is jeopardized [22].

2.2 Demand Response

As mentioned before, DR is defined as a change in the end-user's energy use from a normal consumption pattern in response to a price change or an incentive payment. The implementation of these time-based rates is designed to induce a lower electricity consumption during periods of high wholesale market prices or to ensure the system reliability [23]. The main goal of a DR program is to improve the system's energy efficiency by exploiting the end-users' flexibility. As a result, some of the loads will be shifting from the on-peak to the off-peak periods. In addition, not only do the DR executors make benefits but also the customers obtain monetary compensations for their active participation [24].

From the smart grid perspective, DR is an effective means to reduce high operating costs from expensive generators and mitigate the long-term capacity addition [25]. Furthermore, the implementation of DR can also allow the higher integration of renewable energy using to the power grid and facilitate the utilization of intermittent energy resources in a higher proportion [26] [27]. As a result, the use of DR strategies will allow achieving a more reliable power system. At the same time, the deployment of the electricity market can be enhanced in terms of transparency and efficiency and reaches mutual financial benefits for both the power system operator and all users. Finally, during this process of achieving more efficient uses of the power grid capacity, the emission generation will be reduced, and the environmental impacts will be alleviated.

DR programs use tariffs to incentivize users to modify their usual energy consumption pattern. This enables the opportunity for the power system operator to make an indirect control over end-users' demand. Taking into account the nature of the tariff offered in the DR program, it is possible to divide them into two main programs called incentive-based and price-based, as presented in figure 2.2.

2.2.1 Incentive-based programs

The incentive-based DR programs offer payments to users to motivate demand participation in balancing the generation-consumption imbalance, especially during peak load periods or during system contingencies. The program provides a load modification

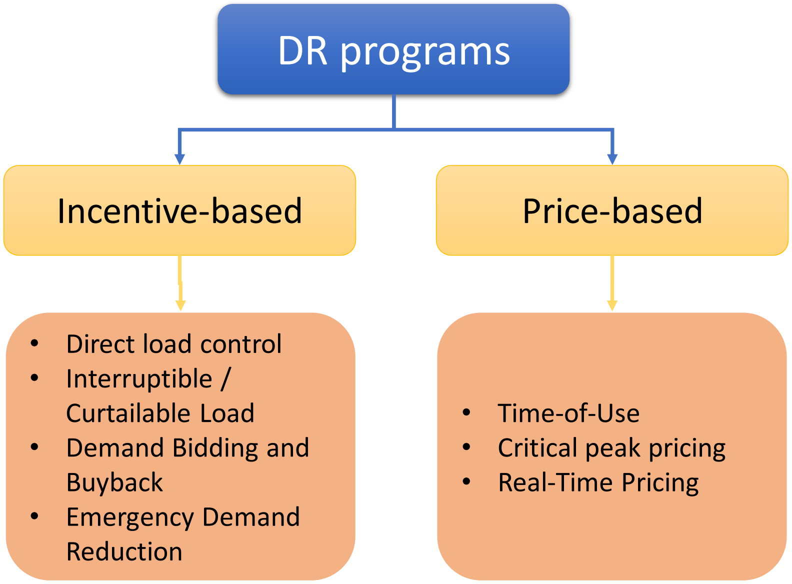


Figure 2.2 Main types of demand response programs, [24].

incentive to the users, resulting in a monetary stimulus separated from electricity prices [28]. The most common incentive programs are:

- **Direct load control:** The system operator has direct access to control specific loads or even manage an end user's entire demand, and in return, incentives are offered to users for their participation [29].
- **Interruptible/Curtailable load:** The users agree to cut down a portion of their interruptible/curtailable load during system contingencies, in exchange of certain incentive discount on electricity bills. In case the users do not curtail, they can be penalized [30].
- **Demand bidding and buyback:** A procedure for peak demand reduction that encourages the large users (1 MW or more) to curtail their load in peak hours, and in return, gain cost saving through rewards. For small users, they will need a third party or an agent to unite and represent them during the bidding process [31].

- **Emergency demand reduction:** During emergency reliability accidents when the grid is out of reserves, the users are notified on very short notice to reduce their demand, receiving incentive payments in return. Through this program, larger users can provide auxiliary services to the power utility, behaving as virtual spinning reserves [32].

2.2.2 Price-based programs

Price-based DR programs provide an alternative to the traditional flat tariffs, evolving to a concept called smart pricing. These schemes are already applied in various countries to a large number of residential customers. For instance, in Ontario, Canada, different price profiles are defined on a seasonal basis (different in summer and winter). Price-based programs define different electricity tariffs at different times. The users will naturally react to this information to avoid bill increases or to obtain benefits. As a result, the consumers will decrease their energy consumption during high price periods and thus reduce their demand at peak hours. This means that smart pricing allows the opportunity of indirect control over users' demand, instead of directly controlling their loads [33].

- **Time-of-use:** The electricity price is defined according to the energy consumption at different time intervals of a day, or different seasons of a year. Typically, each time period for the price is longer than one hour and is defined as on-peak, mid-peak, and off-peak time block. To make the users shift their loads from the on-peak to the off-peak and mid-peak periods, the electricity price is much higher for the on-peak block. The pricing profile for time-of-use (ToU) programs is usually delivered well in advance, and they typically keep unchanged for an extended period [34].
- **Critical peak pricing:** The base structure for this tariff is ToU pricing except

for the days when utilities observe or anticipate high wholesale market prices or power system emergency conditions, that jeopardize the grid reliability. Thus, the normal peak price is substituted by a predefined higher rate to further reduce users' demand. The time and duration of the peak price are predetermined, but the event days are not predetermined [35].

- **Real-time pricing:** Also called dynamic pricing schemes, define electricity prices that vary during the day. Typically the prices change during time intervals of 15 mins or each hour, and they can be deployed on an hour-ahead, or day-ahead basis [36].

2.3 Demand Response Aggregator

DRAs are entities that are allowed to participate in some electricity markets. They are capable of acting as a third-party intermediate between the market and the different actors of the power grid [10]. These aggregators can capitalize on the customers' ability to manage their loads and energy generation. This capitalization is done by sending signals to historically static consumers and then taking their total accumulative capacity to comply with requirements for entering into electricity markets [37].

The role of the DRA faces two crucial challenges. The first one is at the customers level, where the DRA seeks to define DR programs that minimize costs while accurately modeling the customers' behavior in response to economic incentives. The second is in the wholesale market, as the DRA has the challenge of determining the optimal trading options. However, it may face different options ranging from the pool market, where prices are uncertain, to bilateral forward contracts where prices are fixed for a given period [38].

The DRA's ability to captivate the available energy capacities from grid entities is beneficial for operators that need to secure extra system capacity. This allows the power grid to gain advantages of its increasing levels of renewable energy generation from entities that, as singular parts, may not be considered sufficiently valuable to enter into the market [39].

2.4 Optimal Transactive Policy Generation

According to the literature, one of the most exciting topics discussed is the use of control algorithms to facilitate the implementation of DR programs in the residential sector [40]. The objectives of DR optimization algorithms are to further support the DR programs and expand the propagation of DR programs in the energy system. The development of these research works is focused on three different points, the buildings, the market, and the improvement of attributes of the load curve [41]. In order to improve the quality of the load curve, one of the most important approaches is the generation of optimal transactional policies.

Various research studies have examined the potential benefits of DR programs from different perspectives. They have conducted several investigations on the most effective pricing strategies to overcome challenges related to quantifying prices and defining time blocks. The literature has proposed different methods to deal with the generation of optimal price policies. Decentralized methods have been considered to address this issue. For instance, in [42], authors developed a coordination method based on a dynamic pricing strategy to reduce residential bills and aggregated peak load in a day-ahead market. Similarly, in [43], authors used genetic optimization and rolling-horizon algorithms to propose a Time of Use (ToU) pricing strategy and an incentive-based energy management technique, which was employed to decrease the electricity bill and promote the use of renewable energy. The authors in [44] applied dynamic pricing to a day-ahead decentralized coordination problem aimed at reducing the electricity bill through energy sharing and appliance scheduling. Lastly, in [45], authors used a proximal decomposition-based dynamic pricing method to minimize the square Euclidean distance between instantaneous and average energy demand while preserving users' privacy by leveraging a sharing-the-cost mechanism.

While decentralized pricing methods improve the operational performance of electrical grids, require a reliable communication system. As a result, centralized approaches to demand-side services are being promoted as they alleviate the impact of communication failures and provide economically efficient solutions. The utilization of DRAs can provide agents with inexpensive computing equipment to process simple control signals. Centralized strategies for generating optimal price policies have been developed using three main algorithmic mechanisms: game theory, constrained optimization, and RL. Game theory is one of the most commonly used approaches for this purpose. For instance, the authors of [46] have used cooperative game theory to model ToU pricing, while [47] has employed a trilayer Stackelberg game to determine optimal ToU tariffs for a typical community microgrid with prosumers. Authors in [48] have proposed a scalable, hierarchical, transactional approach integrating batteries and model-free control mechanisms and used the Stackelberg game to model negotiations between the distribution system operator and a load aggregator responsible for the coordination and aggregation of a large number of buildings with flexible energy demand. In [49], the same theory was used to characterize the transactive price signal of a DRA based on the Nash equilibrium of the transactive energy in a non-cooperative game. Overall, due to the hierarchical relationship between the players in a DR program, the Stackelberg leadership model is a popular game type widely used for price-based DR studies.

Considering optimization methods for generating transactive policies, authors in [14], propose a profit maximization algorithm that determines optimal prices for an electric utility, considering market constraints. The algorithm's solution is then utilized in a hybrid model that takes into account customers' demand based on their response to the generated price signals. Another study, [15], divides consumers into low and high-energy users and employs a bi-level optimization problem to establish a fair pricing system. This system aims to prevent unfair billing for customers with low energy demand through the demand response decision-making process. The authors of [50] and [51] considered this problem in the policy generation process. The former proposed personalized real-time pricing structures, and the latter implemented a load-based clustering method. These endeavors strive to fulfill users' requirements while ensuring a reliable power supply during peak demand. However, the computation costs associated with processing multiple price policies pose a challenge, potentially hindering real-time applications.

To effectively implement the aforementioned methods, customers must furnish specific information such as initial consumption and satisfaction rate to handle the inherent uncertainty of DR. Previous studies have operated under the assumption that user information is accessible to generate optimal price policies. However, this approach risks customer privacy and may result in a loss of interest in the proposed price policies. Furthermore, it opens the door for dishonest behavior and information concealment, which can undermine the performance of DR programs. For instance, in references [34], [52], and [53], customers are expected to reveal their price elasticity, demand characteristics related to energy conversion and storage devices, and models of responsive loads, respectively. In response, some studies have been conducted considering the significant involvement of customers in developing pricing strategies that minimize their reliance on extensive information disclosure. Reference [10] introduces a pricing scheme based on a non-cooperative scenario, which achieves peak demand reduction for diverse participants with minimal communication requirements. Nevertheless, this approach still necessitates customers to report their total energy consumption, which can impact the accuracy and truthfulness of the proposed mechanism. Moreover, the authors did not clearly define the service provider's objectives in generating price signals, potentially affecting the scalability of their method.

2.4.1 RL-based Transactive Policy Generation

Due to its ability to handle information limitations and load uncertainties, the RL methods have emerged as a promising choice for DR applications due to their ability to deal with uncertainties [54]. In fact, these machine learning approaches are known for their ability to solve problems with hidden information, and their applications in DR programs range from managing household scheduling [55,56] to obtaining near to-optimal transactive policies [57], as they are presented in [58]. However, the learning process in RL algorithms involves a time-consuming training process which results in a significant restriction in the real-world application of RL-based developments.

Concerning implementing RL mechanisms in DR programs, reference [2] presents a deep RL approach to derive optimal incentive policies for incentive-based DR programs. Similarly, in reference [27], deep RL methods are utilized in continuous action domains to address load frequency control challenges arising from renewable energy uncertainties. The authors of reference [16] develop an RL-based decision-making system to assist end-users in selecting the most advantageous Time of Use (ToU) tariffs and monthly rates, thereby minimizing electricity costs and dissatisfaction. Further exploration of RL algorithms in power and energy systems can be found in reference [28], where the focus is optimizing transactive policies in price-based DR programs.

In the context of Real-Time Pricing schemes, references [1] and [29] employ the Q-Learning algorithm, a model-free RL technique, to minimize customer costs. The former considers aggregator profit, while the latter addresses utility cost. These studies utilize information on user dissatisfaction caused by demand reduction to determine the Real-Time Pricing policy. However, their methods require a large number of episodes to converge, making real-world implementations challenging. In reference [30], the authors propose a Monte-Carlo RL technique to optimize retail prices in local micro-grids for a distribution system operator while also protecting end-user privacy. Their approach minimizes the peak-to-average ratio and maximizes profit through energy sales, effectively handling uncertainty problems. However, the assumption that consumer agents are reactive eliminates the negotiation process and disregards the possibility of agents exploring alternative strategies to optimize their consumption. This assumption not only impacts the scalability of the proposed RL-based method for practical applications. These limitations highlight the need for further research in the price policy generation process of DR programs.

2.4.2 Multi-aggregator systems

The literature demonstrates that the policy generation problem has been addressed for different types of DR programs, from incentive-based to price-based [59]. These approaches aim to solve a local problem, assuming the local solutions will lead to a good global solution. 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. Although the proposed approaches have made it possible to identify strategies for the generation of a 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 [60].

The additional effort required for system operators to develop personalized price profiles while considering residents' consumption patterns and preferences becomes increasingly burdensome. However, the concept of DRAs effectively addresses this challenge by facilitating customer participation in a customer-oriented manner, as mentioned in [10]. Despite this, there is limited research on multi-aggregator systems, with only a few works implementing multiagent systems to tackle this issue. In one such work [18], a hierarchical alternating direction method of multipliers (H-ADMM) mechanism is employed to determine load following signals for multiple aggregators. However, this approach assumes that aggregators have direct control over individual devices, potentially compromising customer privacy. Another proposal, presented in [61], suggests a bargaining-based cooperative game to resolve conflicting incentive pricing strategies among multiple aggregators. However, this solution relies excessively on user involvement, which can have drawbacks.

2.5 Synthesis of the literature review

Generally speaking, different approaches have been used in the literature to generate optimal policies for DR programs. All these methods are presented in figure 2.3, as well as the challenges considered in the literature during the transaction policy generation process. From this point, this list of challenges presented, this figure highlights the four target points of our study. However, not all of these approaches are suitable for DRAs, in which these entities act as leaders in the process of exploiting customer flexibility. Considering this third-party player, figure 2.4 provides information on the most common methods applied

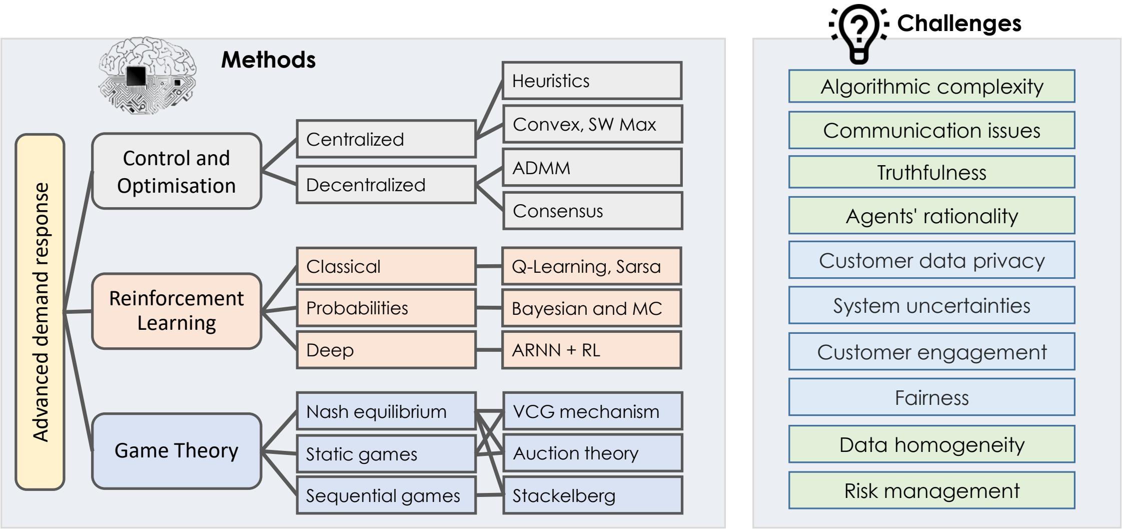


Figure 2.3 Popular methods applied for optimal transactive policy generation.

in the literature in the presence of DRA agents. Many research studies have been done applying game theory, with the Nash equilibrium and Stackelberg games as the preferred. In order to consider the DRA as a leader in the DR program, the Stackelberg games fit very well as they frame the relationship between the aggregator and the residential customers in a leader-follower architecture [48, 49]. In addition to game theory, optimization problems have been considered targeting costs, energy consumption, or welfare [62, 63]. However, to apply the mentioned methods, authors suppose they have access to a large amount of information from residential customers or even direct load control. As a result, it affects users' privacy and makes them refuse their participation in the DR program. Furthermore, it allows the customers to gain advantages of this situation by providing dishonest reports to cheat the system [64]. Finally, RL mechanisms appear to be increasing in popularity due to their ability to deal with system uncertainties and as a valuable option for maintaining customer data privacy.

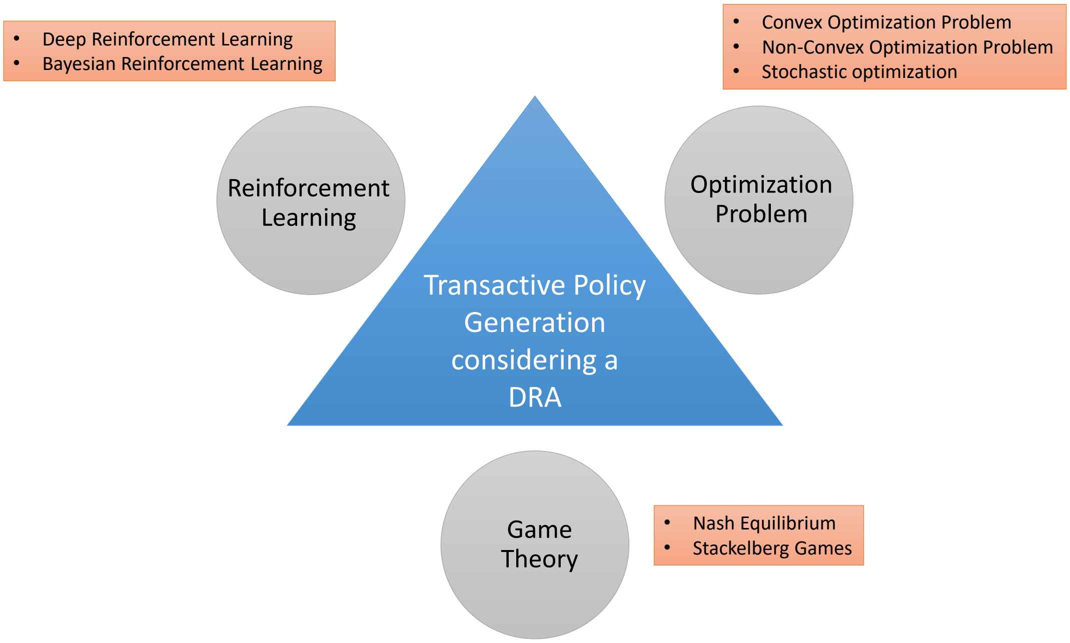

Deep Reinforcement Learning
Bayesian Reinforcement Learning
Reinforcement Learning
Convex Optimization Problem
Non-Convex Optimization Problem
Stochastic optimization
Optimization Problem
Transactive Policy Generation considering a DRA
Game Theory
Nash Equilibrium
Stackelberg Games

Figure 2.4 Methods applied for transactive policy generation considering DRA.

Chapter 3 - Article-based statement of the results

The results of the proposed methodology to achieve the objectives of this research project have been separated into three publications. First, a pricing mechanism is developed to evaluate the architecture of the multi-agent system to ensure convergence, then a price generator function is proposed to parameterize the price policy generation considering constraints from both the market and the supply, and a multi-aggregator architecture is established using the proposed multi-agent architecture and price generator function. The publication status of the articles presented below are as follows:

1. The first article was published in IEEE Access on May 17, 2022.
2. The second was published in Smart Energy from Elsevier on March 27, 2024.
3. Finally, the third manuscript was submitted to Sustainable Energy, Grids and Networks from Elsevier on April 9, 2024, and is currently under revision.

3.1 Multi-agent architecture for transaction policy generation

3.1.1 Background

In this first part, we consider the problem of developing a multi-agent system for the interaction between the residential agents and a DRA agent as market players. The interaction between these entities will allow the DRA to optimize the price policy generation process. So, automated residential agents need to be constructed to ensure rational responses to DRA agent actions. For the DRA agent case, as the only source of information for him will be the DR to respect customers' privacy, a data-driven mechanism based on RL will be developed to define near-to-optimal price policies.

The DR program established by the DRA will be a discount-based ToU electricity pricing strategy. This agent will offer discounts at different hours of the day to encourage users to change their consumption patterns. The decision-making process is done by applying a deep RL technique to deal with the different sources of uncertainty due to the lack of information from the demand side. To guarantee an improvement in the quality of the aggregated load profile, the DRA agent will try to maximize the load factor (inverse of the PAR) while, at the same time, maximizing its profit during the definition of the different discount rates. Using the load factor in the objective function does not admit the use of gradient-based techniques for the optimization process, making RL a more valuable approach.

For the residential agent, a thermal model is constructed for the space heating system by means of historical consumption data from different Quebec houses located in Trois-Rivieres. This model will enable heating to be used as a source of flexibility by the users. The HEMS will determine its consumption plan in response to the price policy by solving an optimization problem utilizing model predictive control. The goal of the residential agent is to minimize its bill for consuming energy while maintaining the thermal comfort of the customers. Finally, a regularization mechanism of residential agent response is applied to guarantee convergence of the multi-agent system, based on a proximal decomposition approach.

3.1.2 Methodology

The reinforcement learning environment is composed of a set of twenty residential houses. To determine the customers' thermal preferences, the information is obtained from a previous work conducted in Quebec's context by [65]. By exploiting the collected historical data, the parameters of the state-space representation for the thermal model are estimated utilizing a ridge regression mechanism [66]. Finally, a statistical data generation process is used for the non-controllable loads, which are added to the thermal model output to construct the overall power profile of each house.

Once the residential agents are ready to respond to price signals, a historical day is selected for the offline training of the RL-based DRA. The following analyses were done to ensure the best performance of the proposed price-based DR program:

1. A performance comparison of different Deep RL techniques, which allow the selection of PPO as the target mechanism for the implementation of the DR program.
2. A multi-agent system convergence analysis based on the selection of the regularization parameter τ for the residential agents.
3. A comparative study between the RL-based proposed approach and a Proximal decomposition mechanism from the literature in terms of load factor improvement and DRA's profit.

Once these evaluations were done offline, the performance of the DRA decision-making process was evaluated online for consecutive days by randomly selecting different external temperature profiles from the database. For this purpose, the DRA agent is trained on a historic winter day. Once the agent learns how to generate the price-based policies, the online evaluation process starts for consecutive winter days, selecting daily winter profiles randomly. The proposed procedure is summarized in Figure 3.1.

3.1.3 Outcomes

This work proposes generating a discount-based ToU tariff as a valuable option for encouraging users to participate in a DR program. The presented approach developed a data-driven DRA for generating near-to-optimal hourly ToU tariffs. This mechanism offers a DR service to the grid operator to exploit the flexibility potentials from the demand side

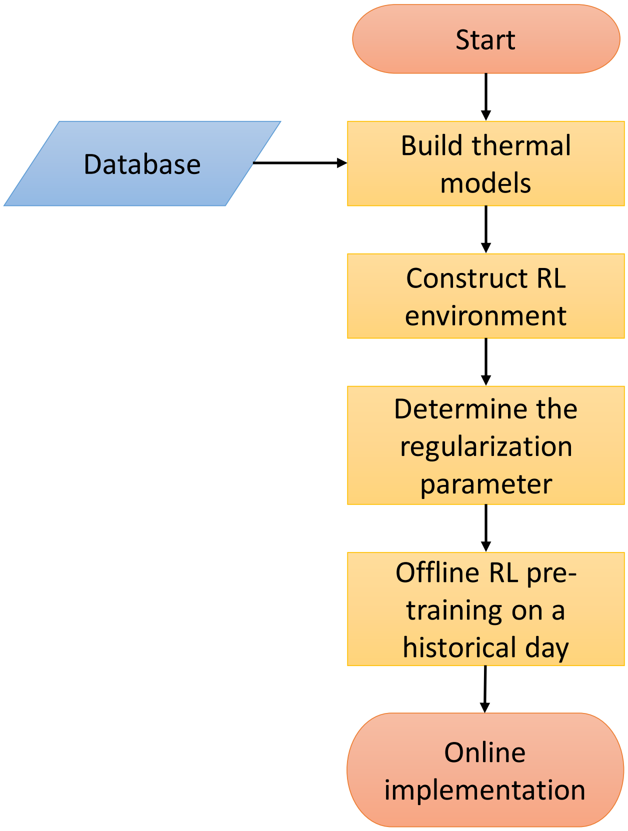

Start
Database
Build thermal models
Construct RL environment
Determine the regularization parameter
Offline RL pre-training on a historical day
Online implementation

Figure 3.1 Block diagram for the proposed procedure.

to help with peak shaving needs. In this process, the DRA agent determines price policies based on discounts captured by minimal information exchange with end-user agents, as the only information needed for determining the price policies is the DR. This design allows the reduction of infrastructural needs for communication and maintains customer agents' privacy within reliable interactions.

In terms of implementation, this study has recommended an RL algorithm for constructing a promising DR system. Furthermore, the proposed approach provides an offline training phase strategy to deal with the time-consuming convergence of RL techniques. This proposition reduced the convergence time from more than 1000 days to less than 20 days, enabling online implementations. Regarding optimization, the aggregator agent realizes a trade-off between load factor and total revenue as two contrary objectives since the DR depends on the economic incentive offered, i.e., the greater the incentive, the greater the expected exploited flexibility. The obtained results were compared with two common RL techniques where the proposed RL mechanism manifests the superior performance of the recommended structure through high and fast convergence rates. And a proximal decomposition-based coordination scheme is compared as well where the RL-based DRA can achieve a lower reduction of its profit, although the near-to-optimal tariff is based on discounts. On the other hand, a larger income reduction based on the proximal decomposition method evidences that the monetary sacrifice in a DR program can be high if it is not controlled. These results highlight the suggested mechanism's efficiency.

IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received April 29, 2022, accepted May 12, 2022, date of publication May 17, 2022, date of current version May 25, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3175839

A Discount-Based Time-of-Use Electricity Pricing Strategy for Demand Response With Minimum Information Using Reinforcement Learning

ALEJANDRO FRAIJA[1], KODJO AGBOSSOU[1], (Senior Member, IEEE), NILSON HENAO[1], SOUSSO KELOUWANI[2], (Senior Member, IEEE), MICHAËL FOURNIER[3], AND SAYED SAEED HOSSEINI[1], (Student Member, IEEE)

[1]Department of Electrical and Computer Engineering, Hydrogen Research Institute, University of Quebec at Trois-Rivières, Trois-Rivieres, QC G8Z 4M3, Canada
[2]Department of Mechanical Engineering, Hydrogen Research Institute, University of Quebec at Trois-Rivières, Trois-Rivieres, QC G8Z 4M3, Canada
[3]Laboratoire des Technologies de l'Energie, IREQ, Shawinigan, QC G9N 0C5, Canada

Corresponding author: Alejandro Fraija (alejandro.jose.fraija.ochoa@uqtr.ca)

This work was supported in part by the Laboratoire des Technologies de l'Énergie d'Hydro-Québec, in part by the Natural Science and Engineering Research Council of Canada, and in part by the Foundation of Université du Québec à Trois-Rivières.

ABSTRACT Demand Response (DR) programs show great promise for energy saving and load profile flattening. They bring about an opportunity for indirect control of end-users' demand based on different price policies. However, the difficulty in characterizing the price-responsive behavior of customers is a significant challenge towards an optimal selection of these policies. This paper proposes a Demand Response Aggregator (DRA) for transactive policy generation by combining a Reinforcement Learning (RL) technique on the aggregator side with a convex optimization problem on the customer side. The proposed DRA can maintain users' privacy by exploiting the DR as the only source of information. In addition, it can avoid mistakenly penalizing users by offering price discounts as an incentive to realize a satisfying multi-agent environment. With an ensured convergence, the resultant DRA is capable of learning adaptive Time-of-Use (ToU) tariffs and generating near-to-optimal price policies. Moreover, this study suggests an off-line training procedure that can deal with issues related to the convergence time of RL algorithms. The suggested process can notably expedite the DRA convergence and, in turn, enable online applications. The developed method is applied to a set of residential agents in order to benefit them by regulating their thermal loads according to generated price policies. The efficiency of the proposed approach is thoroughly evaluated from the standpoint of the aggregator and customers in terms of load shifting and comfort maintenance, respectively. Besides, the superior performance of the selected RL method is represented through a comparative study. An additional assessment is also conducted by use of a coordination algorithm to validate the competitiveness of the recommended DR program. The multifaceted evaluation demonstrates that the designed scheme can significantly improve the quality of the aggregated load profile with a low reduction in the aggregator's income.

INDEX TERMS Demand response, demand response aggregator, time-of-use tariffs, reinforcement learning.

NOMENCLATURE

Indices

t	Iteration index.
i	House index.
k	Time-step index.

Parameters

ω	Trade-off weighting factor of the reward function.
τ	Regularization parameter of the proximal decomposition method.
$x^i_{\min}$	Lower bound of i^{th} household internal temperature.
$x^i_{\max}$	Upper bound of i^{th} household internal temperature.

The associate editor coordinating the review of this manuscript and approving it for publication was Inam Nutkani.

$u_{\max}^{i,Th}$	Heating system capacity of i^{th} house at time-step k.

Variables

s_t	State at episode t.
a_t	Action at episode t.
μ_t^h	Normalized hourly average of the aggregated energy consumption.
$\bar{u}^h$	Average energy consumption at hour h.
α_t^h	Normalized energy price at hour h.
λ_t^h	Energy price value at hour h.
ξ	Initial flat energy price.
u_k^i	Energy consumption of i^{th} house at time-step k.
$u_k^{i,Th}$	Thermal energy consumption of i^{th} house at time-step k.
$u_k^{i,NC}$	Energy consumption of non-controllable loads of i^{th} house at time-step k.
x_k^i	Indoor temperature of i^{th} house at time-step k.
w_k^i	Outdoor temperature at time-step k.
δ_k^i	Thermal discomfort factor of i^{th} house.
x_{sp}^i	Set-point temperature profile of i^{th} house.

Functions

R_t	Reward function at episode t.
$\hat{A}_t$	Advantage at episode t.
LF	Load factor of the aggregated energy consumption profile.
Pr	Aggregator's income sacrifice ratio.
$TC(u_k^{i,Th})$	Thermal comfort function.

Abbreviations

DR	Demand Response.
DRA	Demand Response Aggregator.
RL	Reinforcement Learning.
ToU	Time-of-Use.
DB-ToU	Discount Based Time-of-Use.
MDP	Markov Decision Process.
PPO	Proximal Policy Optimization.
ESH	Electric Space Heating.

I. INTRODUCTION

The rapid increase in energy needs and associated greenhouse gas emissions has created significant challenges to traditional power systems. This issue can be relieved by the promise of smart grids that bring about a modern power system with efficient alternatives regarding the energy transition concept [1]. In the context of the smart grid, Demand Response (DR) is favored as an effective mechanism to mitigate peak demand by utilizing communication technologies and advanced metering infrastructures. DR programs employ price and incentive signals to change end-users' consumption patterns, provide stability, balance energy resources, and bring economic efficiency to grid stakeholders [2], [3]. DR programs devise various pricing strategies for alleviating daily peak load. These schemes aim to shift energy consumption from on-peak to off-peak hours. The main idea is to define higher price rates for on-peak hours so that users shift their load in order to avoid extra electricity bills. However, users' response can result in generating new peaks since it increases energy demand during off-peak hours [4]. This issue can result from DR methods based on traditional flat-rate electricity tariffs. Accordingly, other pricing strategies have been proposed to provide alternatives to former policies. These techniques generally offer price-based DR programs in which utilities or aggregators are in charge of recommended policies considering the historical behavior of end-users' load profiles [5]. They include Real-Time Pricing (RTP), Time-of-Use (ToU) pricing, and Critical Peak Pricing (CPP), where RTP and ToU are the most commonly used means [6]. RTP is a scheme in which the electricity price varies over short periods, normally hourly, with regard to the real-time production cost. On the other hand, ToU pricing is a tariff in which constant electricity prices are considered for lengthy time intervals, typically hours of the day or days of the week [7]. The latter is normally preferred by both grid operators and customers, and, thus, has been the main focus of the relevant literature [4].

A. RELATED WORK

Research works have explored DR programs from different aspects to reveal their potential benefits. They have carried out various studies on optimal pricing strategies to overcome the challenges related to price quantification and time blocks definition [6]. Particularly, different approaches have been proposed in the literature to deal with optimal price policy generation. From one side, decentralized methods have been considered to address this matter. Authors in [8] have developed a coordination method based on a dynamic pricing strategy to reduce the residential bill and aggregated peak load in a day-ahead market. In [9], the authors have proposed a ToU pricing strategy and an incentive-based energy management technique by means of genetic optimization and rolling-horizon algorithms. They have employed this framework to decrease the electricity bill and increase the use of renewable energy. The authors in [10] have applied dynamic pricing to a day-ahead decentralized coordination problem. Their strategy has been aimed at reducing the electricity bill through energy sharing and appliance scheduling. In [11], the authors have developed a proximal decomposition-based dynamic pricing method to minimize the square Euclidean distance between instantaneous and average energy demand. In addition, they have exploited a sharing-the-cost mechanism while preserving the privacy of users.

Although decentralized pricing methods can improve the operational performance of electrical grids, they require a reliable communication system. Accordingly, centralized approaches to demand-side services are promoted. Centralizing pricing tariffs not only alleviate the impact of communication failures but also provide economically efficient solutions. In this context, agents can use inexpensive computing equipment to process simple control signals (policies), offered by a DR Aggregator (DRA) [12]. Centralized strategies for generating optimal price policies have been carried out in the literature based on three main algorithmic mechanisms comprising game theory, constrained optimization, and Reinforcement Learning (RL). The game theory is one of the most utilized approaches for this purpose. The authors in [7] have employed a cooperative game theory to model ToU pricing. In [13], a trilayer Stackelberg game has been exploited to determine optimal ToU tariffs for a typical community microgrid with prosumers. In [14], the authors have proposed a scalable, hierarchical, transactional approach to integrate batteries and model-free control mechanisms. They have used the Stackelberg game to model negotiations between the distribution system operator and a load aggregator responsible for efficient coordination and aggregation of a large number of buildings with flexible energy demand. The authors in [15] have utilized the same theory to characterize the transactive price signal of a DRA based on the Nash equilibrium of the transactive energy in a non-cooperative game. It is worth mentioning that the Stackelberg leadership model is a popular type of game that has been widely used for ToU-based DR studies.

In addition to the game theory, the problem of transactive policy generation has been tackled by optimization methods. In [16], a profit maximization algorithm has been proposed to accomplish optimal prices for an electric utility under market constraints. The optimal solution has been adopted for a hybrid model of customers' demand according to their response to generated price signals. In [17], consumers have been categorized into low and high energy users. Consequently, a bi-level optimization problem has been implemented to realize a fair pricing system. This mechanism has been intended to deal with the possibility of unfair billing to customers with low energy demand through the DR decision-making procedure. In this regard, it has carried out an individual billing strategy for every detected homogeneous consumer. The same issue has been encountered by the authors in [18] and [19]. In order to avoid imposing an unfair penalty, the former has developed a personalized real-time pricing structure while the latter has employed a load-based clustering manner. As a result, these works have attempted to meet users' desires while maintaining a reliable power supply during peak demand. Nevertheless, they have undergone notable computational costs due to processing multiple price policies, which can hinder real-time applications.

A fruitful application of the above methods needs customers to provide specific information such as initial consumption and satisfaction rate to handle the inherent uncertainty of DR programs. Therefore, the previous studies have assumed that users' information is accessible in order to generate optimal price policies. However, such reliance upon customers can jeopardize their privacy and cause them to lose interest in generated price policies. Conversely, it can create opportunities for hiding information and interacting in a dishonest manner, which can, in turn, reduce the performance of DR programs. This matter can be specifically exemplified by the proposed methods in [20], [21], and [22]. In [20], the authors have developed an optimal ToU pricing strategy in which consumers' price elasticity must be known. In [21], the authors have practiced a similar procedure in which customers' demand properties related to energy conversion and storage devices are required. In [22], the authors have executed a minimization problem in which the objective function must be provided by the model of responsive loads. The challenges caused by users' excessive involvement have stimulated the development of pricing strategies that reduce the need for their information. In [23], the authors have proposed a pricing scheme with minimal communication requirements based on a non-cooperative scenario. They have proved the existence of a Nash equilibrium to achieve peak demand reduction for heterogeneous players with minimum interactions. Nevertheless, their proposed approach requires customers to report their total energy consumption within every game period. Subsequently, their solution to the problem can be significantly affected by the accuracy and truthfulness of the provided information. Besides, they have not clearly defined the objectives of the service provider for generating price signals, which can affect the scalability of their method.

Recently, the RL method has become a viable option for DR exercises due to its ability to deal with both information limitations and load uncertainties. In fact, this machine learning technique is known for its capability to solve problems with hidden information. In [24] and [25], RL methods have been utilized to manage household load scheduling. In [2], a deep RL approach has been implemented to obtain optimal incentive policies through an incentive-based DR program. Likewise, the authors in [26] have applied deep RL methods in continuous action domains for load frequency control against renewable energy uncertainties. In [27], the authors have constructed an RL-based decision-making system to assist end-users with selecting the most beneficial ToU tariffs and monthly rates and, consequently, minimizing their electricity and dissatisfaction costs. Different applications of RL algorithms in power and energy systems can be studied in [28]. Particularly, RL techniques have been used to attain optimal transactive policies in price-based DR programs. The authors in [3] and [29] have employed the Q-Learning algorithm, as a model-free RL, for RTP schemes. While both studies have aimed to minimize the customer cost, the former has considered the aggregator profit, and the latter has dealt with the utility cost. They have exploited information about user dissatisfaction because of demand reduction to determine the RTP policy. However, their methods involve running thousands of episodes to reach a convergence point, which makes real-world implementations difficult. In [30], the authors have developed a Monte-Carlo RL technique to optimize retail prices in local micro-grids for a distribution system operator while protecting end-user privacy. Their method has allowed for minimizing the peak-to-average ratio and maximizing the profit by selling energy. Additionally, their RL approach can handle the intractability of the problem under a great deal of uncertainty. However, it eliminates the negotiation process since it assumes that consumer agents are reactive. This assumption rules out the fact that the agents can be proactive and explore other strategies to optimize their consumption. In addition, it affects the scalability of their proposed RL-based method for relevant applications. Besides, they have not elaborated on the convergence time as a critical factor in implementing RL methods while reporting the results. Indeed, the above restrictions necessitate further investigations into the price policy generation procedure of DR programs.

B. MOTIVATION AND CONTRIBUTION

Inspired by the previous works, this paper seeks to overcome practical difficulties in achieving optimal price policies. From one side, it deals with the possibility of mistakenly penalizing users within the price generation process through a computationally efficient mechanism. From the other side, it handles the concerns related to users' privacy and interaction with the aggregator by completely avoiding the utilization of their information. In fact, overlooking these issues can violate customers' satisfaction and decline their participation in DR programs. As a result, this study makes the following contributions.

1) It proposes a DRA that is able to avoid penalizing users by generating Discount Based ToU (DB-ToU) tariffs. The proposed DRA takes advantage of discounts as an incentive for residential users to exploit their demand flexibility and, consequently, flatten their aggregated power consumption.
2) It develops a procedure that can generate near-to-optimal price policies with no access to end-user internal information. The designed DRA is able to learn customers' behavior towards energy usage only by utilizing their response to transactive policies and handling uncertainties related to the lack of domestic information, which varies from user to user.
3) It constructs a multi-agent environment with ensured convergence by combining an RL method on the aggregator side with an optimization problem on the customer side. Most importantly, the suggested DRA adopts a pre-training strategy that remarkably decreases the convergence time of the RL algorithm and improves its online performance.

The rest of the paper is organized as follows: Section II presents the methodology for formulating the proposed DRA. Section III provides the results and discussion, followed by concluding remarks in Section IV.

II. METHODOLOGY

In a residential distribution grid, operated by automated agents, a DRA is in charge of managing the load flexibility of a group of residences [31]. It provides transactive policies to motivate customers to change their energy consumption and consequently improves the quality of the aggregated load profile. The proposed mechanism targets a group of residential buildings, equipped with energy-intensive controllable loads. Fig. 1 illustrates the methodology for the proposed DR program. In this procedure, the aggregator agent is running a day-ahead pricing scheme. It communicates price signals to residential agents in order to decrease peak load according to their response. To be specific, it offers price discounts during different hours of the day to manage the aggregated demand. At the end of the day, the DRA amasses consumption profiles, calculates rewards, and generates the next policy. In the following, the reinforcement learning method and the reward mechanism for the DRA and residential agents are detailed.

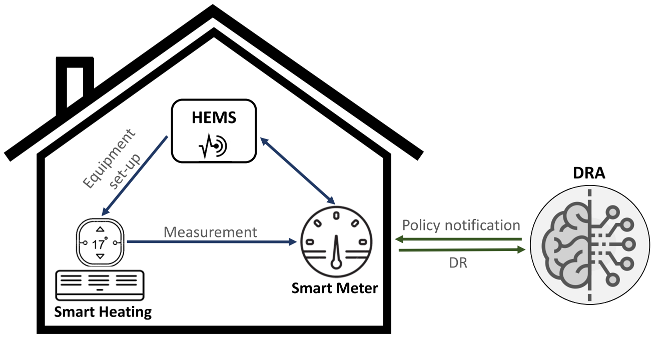


FIGURE 1. Automatic energy management under the price-based DR program.

A. REINFORCEMENT LEARNING

The targeted scenario considers a multi-agent system that is composed of a set of residential agents and a DRA agent. The aggregator agent is an RL agent that executes a trial-error process to learn from an environment as part of a DR program. Generally, this agent chooses actions according to a given state and receives rewards through interacting with the environment [32]. The interaction between the aggregator agent and the environment is represented as a Markov Decision Process (MDP) and is characterized by

1) State, s_t, that presents the hourly average of the aggregated energy consumption,
2) Action, a_t, that explains the established ToU price policy,
3) Reward, R, that predicts the DRA profit according to a chosen action through $R_s^a = \mathbb{E}[R_{t+1}|s_t = s, a_t = a]$.
4) And the discount factor, $\gamma \in [0, 1]$, that defines the importance of the future rewards for the current decisions. Higher values of γ expresses that future rewards have a higher impact on the decision making process.

It should be noted that an MDP is an extension of Markov chain in which the future state, s_{t+1} depends only on the current state, s_t and the current action, a_t. Given the component γ, it is possible to calculate the 'return', G_t, as the future discounted reward. In fact, the task of the RL agent is to collect as many high rewards as possible. Accordingly, the discount factor, γ, is used to realize a bounded reward, G_t, in terms of $R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \gamma^3 R_{t+3} + \cdots$ and avoid an unboundedness problem due to a growing sum (infinite case).

The aggregator agent learns the policy, π, by interacting with the environment. This policy fully describes the behavior of the agent and represents a distribution over the pricing actions considering the states [35]. Afterward, the state value function of MDP, $V_\pi(s)$, is determined as the expected return given the starting state and the policy. Additionally, the state-action value function, $Q_\pi(s, a)$, represents the expected return, starting from the state s, taking the action a, and following the policy π [24]. The optimal policy, π^*, results in the optimal state value function, $V^*(s)$. In fact, this function is obtained when the optimal policy is selected by the RL agent [36]. The MDP is solved when the optimal value function is found since it represents the maximum reward for the state s that can be obtained from the system. Similarly, the optimal state-action value function, $Q^*(s, a)$, is realized when the optimal policy is chosen by the RL agent in the state s to have the action a [37]. $Q^*(s, a)$ represents the maximum reward that can be obtained from the state s and the action a.

The proposed approach employs the Proximal Policy Optimization (PPO) as a policy gradient method. The PPO algorithm is used to optimize the policy $\pi_\theta(a, s)$ based on the policy parameter θ. This technique defines a reward function, $J(\theta)$, that depends on $\pi_\theta(a, s)$ and is maximized with respect to θ [32]. PPO is an algorithm with data efficiency and reliable performance, similar to advanced policy gradient methods such as Trust-Region Policy Optimisation (TRPO). These methods try to stabilize agent training by avoiding big policy alterations (updates on θ) per state. However, PPO is a less complex design that takes advantage of first-order techniques instead of complex second-order schemes or hard constraints like KL-divergence [38], [39]. The Algorithm 1 represents the PPO technique for which the objective function, $J(\theta)$, is formulated by,

$$J(\theta) = \hat{\mathbb{E}}_t[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \quad (1)$$

where

- θ is the policy parameter,
- $\hat{E}_t$ is the expectation over episode t,
- $r_t(\theta)$ is the probability ratio between new and old policies as $\pi_\theta(a_t|s_t) \,/\, \pi_{\theta_{old}}(a_t|s_t)$,
- $\hat{A}_t$ is the estimated advantage at episode t as $-V(s_t) + \gamma R_t + \cdots + \gamma^{M-t+1} R_{M-1} + \gamma^{M-t} V(s_M)$ where M is the batch size,
- And ϵ is the hyperparameter for clipping. This parameter avoids large deviations in the updated θ considering θ_{old} by clipping the ratio at the interval $[1-\epsilon, 1+\epsilon]$ [40].

Algorithm 1 PPO Algorithm

```
Input: initial policy parameters $\theta_0$, clipping threshold $\epsilon$, batch size $M$.
for $t = 0, 1, 2, \ldots$ do
    Define the normalized action $a_t$. (Price policy defined by the aggregator agent)
    Get the normalized state $s_t$. (Residential agents' response)
    Calculate the reward $R_t$.
    Collect the set of partial trajectories $\{(s_t, a_t, R_t, s_t + 1)\}$ on policy $\pi_t = \pi(\theta_t)$.
    Estimate advantage $\hat{A}_t$.
    if $t$ mod $M = 0$ then
        Compute policy update
            $\theta_{t+1} = \arg\max_{\theta} \sum_{j=0}^{M} J(\theta)$
        via stochastic gradient ascent with Adam [33].
    end
end
```

B. DEMAND RESPONSE AGGREGATOR

The DRA is in charge of defining the price policy that is applied for the next 24 hours. Each episode, t, starts with sending the price policy and waiting for the response of the residential agents in terms of power demand. The RL aggregator performs an initial offline training by exploiting the information of a specific day. Subsequently, the trained agent is deployed to provide the transactive price signal for the following days. As a result, the DRA learns to carry out near to optimal pricing policies for a given set of houses by using aggregated energy demand data. In this regard, the state $s_t \in S$ in the MDP can be defined as $s_t = \{\mu_t^1, \mu_t^2, \ldots, \mu_t^{24}\}$ where $\mu_t^h = \dfrac{\bar{u}^h}{\max\limits_{h\in\{1,\ldots,24\}} \{\bar{u}^h\}}$ is the normalized average of the aggregated consumption $\bar{u}^h$ at hour h. The agent selects a normalized action $a_t \in A$ as $a_t = \{\alpha_t^1, \alpha_t^2, \ldots, \alpha_t^{24}\}$. Considering that ξ is the initial price policy, applied to the power grid, and λ_t^h is the price, decided by the DRA for the next hour, the price value at each hour, h, of the day is calculated using α_t^h through,

$$\lambda_t^h = \xi \alpha_t^h \quad (2)$$

Accordingly, a price constraint based on (3) is established by the DR program.

$$0 \le \alpha_t^h \le 1 \quad \forall h \in \{1, 2, \ldots, 24\} \quad (3)$$

This restriction maintains a generated transactive policy lower than the initial tariff, ξ, by constraining the action space of DRA. As a result, it provides residential agents with $\lambda_t^h \le \xi$. Finally, the reward function considers two main objectives, intended by the agent to maximize. They consist of improving the aggregated load profile quality and achieving the optimal DB-ToU tariff with lower aggregator's income sacrifice. The former is aimed at load factor correction in peak reduction, which is the inverse of the peak-to-average ratio. Considering $\mathbf{u} = \{u_1, u_2, \ldots, u_N\}$ as the overall discretized energy consumption profile, the load factor can be calculated through,

$$LF = \frac{\frac{1}{N}\sum_{k=1}^{N} u_k}{\max_k\{\mathbf{u}\}} \tag{4}$$

Besides, the latter is sought by offering price discounts to the houses for shifting their loads without sacrificing the aggregator's income. To be specific, the aggregator agent defines the optimal policy by comparing DB-ToU and constant bills together. Being $u_{0,k}$ the energy consumption when the price is ξ, this comparison is performed by quantifying the aggregator sacrifice based on the ratio between both bills, computed through,

$$Pr = \frac{\sum_{k=1}^{N} u_k \lambda_k}{\xi \sum_{k=1}^{N} u_{0,k}} \tag{5}$$

According to (4) and (5), the agent reward function at the episode t can be explained by,

$$R_t = \omega LF_t + (1-\omega)Pr_t \tag{6}$$

where ω is a weighting factor that allows a trade-off between the aforementioned objectives. The aggregator agent tries to maximize the return by using the proposed reward function. This non-linear objective function balances load factor and total revenue as two conflicting terms. The RL approach enables the utilization of the proposed reward function in (6) since it is not a differentiable operation that can be optimized through gradient-based methods. Generally, RL methods facilitate executing non-differentiable reward functions on the aggregator side. It should be highlighted that this advantage increases the versatility of the recommended DRA for actual implementations.

On the other hand, the price constraint (3) established by the proposed DB-ToU scheme, always provides participants with benefit. Users are never penalized since they receive the initial price without any discount in the worst scenario. This, in turn, boosts customers' motivation for participating in DR program. In addition, the proposed reward function uses the initial energy consumption $\mathbf{u_0} = \{u_{0,1}, u_{0,2}, \ldots, u_{0,N}\}$ from the constant tariff exercise. This practice provides the aggregator with prior knowledge about users' energy consumption preferences and helps provide useful information about the price responsive behavior of the residential agents.

C. RESIDENTIAL ENVIRONMENT

A case study of residential houses, located in Quebec, Canada, during winter is considered in this work. Buildings in the Quebec region represent a specific example of energy consumption. Due to long cold climates, they consume a massive amount of heating energy, which is mainly supplied by electricity. In this district, Electric Space Heating (ESH) systems account for more than 60% of energy consumption [41]. In this case study, the residential environment is composed of 20 agents. The residential agents are capable of controlling their ESH demand by employing a Model Predictive Control (MPC). To be specific, the MPC is applied to thermal models of houses in order to estimate their indoor temperature on a daily basis [42]. The decision-making process of this model is executed based on the maximization of users' Social Welfare Function. Accordingly, an optimal decision is made by satisfying individual participants' comfort, which is maintaining the temperature setpoint (the reference) while minimizing the energy cost. Therefore, they can take advantage of the price discounts, offered by the DRA. In fact, ESH systems, as thermal loads, can provide residential agents with energy flexibility to modify their demand under the DR program. The total energy consumption of the residential agent i at the time-step k is,

$$u_k^i = u_k^{i,Th} + u_k^{i,NC} \tag{7}$$

where $u_k^{i,Th}$ and $u_k^{i,NC}$ are the energy demand of the thermal and other loads (assumed to be non-controllable), respectively. The dynamic thermal response of the houses is described by the state-space representation model to avoid high computational complexity [43]. For the same agent, i, this linear model computes the future value of indoor temperature, x_{k+1}^i, depending on the current amounts of indoor temperature, x_k^i, outdoor temperature, w_k^i, and ESH demand, $u_k^{i,Th}$, based on,

$$x_{k+1}^i = Ax_k^i + Bw_k^i + Cu_k^{i,Th} \tag{8}$$

where A is the state matrix while B and C are the input matrices associated with the weather and heating sources, respectively. The residential agent controls the thermal loads to minimize the cost of energy consumption considering occupants' desires. Thermal comfort desires are used to formulate the concave utility function through [44],

$$TC(u_k^{i,Th}) = -\delta_k^i(x_{\text{sp}}^i - x_k^i)^2 \tag{9}$$

where for the agent i in (8), x_{sp}^i presents the set-point temperature profile, x_{k}^i represents the internal temperature profile, and δ_k^i is the discomfort factor. This latter element characterizes users' willingness to sacrifice their thermal comfort in order to reduce the bill. To be specific, it defines periods of the day within which the comfort level varies between high and low boundary conditions. In order to perform a realistic scenario, the values of δ_k^i are determined according to the comfort preferences in the Quebec residential sector, presented in [45].

Since the residential agents solve their optimization problem in a selfish way, they do not cooperate with each other. Dealing with the individuals who attempt to maximize their own profit can expose the proposed approach to the *prisoner's dilemma*. In order to address this issue, a proximal decomposition approach is established by penalizing the residential agents' demand modification based on the regularization parameter τ. The penalization is applied to the difference between the current energy consumption at episode t and its previous amount at episode $t-1$. Considering the utility function in (9), the individual welfare can be expressed by,

$$W = \sum_{k=1}^{N} TC(u_k^{i,Th}) - \lambda_k u_k^i - \tau(u_{t,k}^i - u_{t-1,k}^i)^2 \qquad (10)$$

The goal of residential agents is to maximize their individual welfare. As a result, the dual problem of the agents' cost function can be formulated through,

$$\begin{aligned} &\underset{u^i=\{u_k^i\}_{k=1}^{T}}{\text{Minimize}} \sum_{k=1}^{N} \delta_k (x_{\text{sp}}^i - x_k^i)^2 + \lambda_k u_k^i + \tau(u_{t,k}^i - u_{t-1,k}^i)^2 \\ &\text{subject to Eqnarray(8)} \\ &\qquad x_k^i \in [x_{\text{min}}^i, x_{\text{max}}^i] \\ &\qquad u_k^{i,Th} \in [0, u_{\text{max}}^{i,Th}] \\ &\qquad u_k^i = u_k^{i,Th} + u_k^{i,NC} \qquad (11) \end{aligned}$$

where the parameters x_{min}^i and x_{max}^i are the lower and upper bounds of the allowed internal temperature, respectively, and $u_{\text{max}}^{i,Th}$ is the heating system capacity within time slot k. It should be noted that the temperature bounds are set by the user for the thermostat. The equation (11) is a convex optimization problem that is solved by using Disciplined Convex Programming (DCP). The optimal solution is calculated by means of the Embedded Conic Solver (ECOS) through the Python-embedded modeling language for convex optimization, CVXPY.

III. RESULTS AND DISCUSSION

A. REINFORCEMENT LEARNING ENVIRONMENT PREPARATION

The reinforcement learning environment is composed of a set of twenty residential houses. The electric heating system information of these houses is obtained from a previous study, conducted by the authors for the case of Quebec in [46]. This information that comprises simulated ESH demand, as well as internal and external temperatures, is used to create the thermal dynamic response model of the houses, described by (8). For this purpose, the parameters of the state space representation of each house are estimated by means of the Ridge regression technique [47]. Additionally, a data generation process is used to create the non-controllable appliances' load based on the same study [46]. This process employs the power consumption distribution of these devices, captured from actual data of eight houses in Quebec during winter, to generate their demand through a sampling procedure. Subsequently, the ESH and non-controllable loads are added to construct the overall power profile of each house. Finally, the user preference and set-point temperature profiles of the houses are acquired from [45], in which the author has investigated these features in Quebec households. The above practice provides the twenty houses with different electricity consumption patterns, which is pertinent to the Quebec region. Fig. 2 exemplifies the aggregated energy consumption behavior of heating and uncontrollable demand in twenty houses for a typical day in winter 2018.

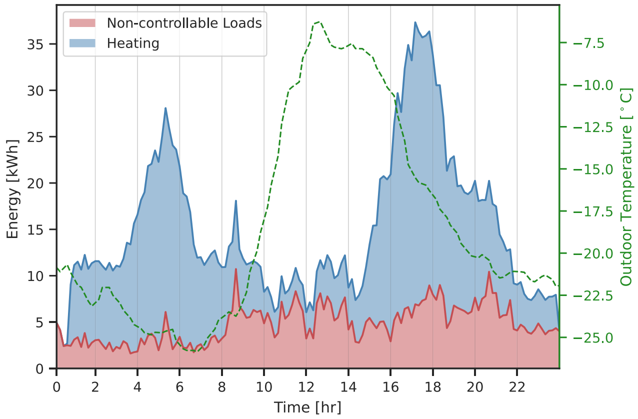


FIGURE 2. Aggregated energy consumption behavior of twenty residential agents in correlation with outside temperature on a typical winter day.

The operation of residential agents in the RL environment is carried out by OpenAI Gym as a toolkit for exploring RL algorithms. In this environment, the aggregator agent starts the pre-training phase on a randomly chosen day. Accordingly, the aggregator learns the optimal DB-ToU price policy for the selected day by applying PPO while taking into account the reward function, presented in (6). Afterwards, it defines the DB-ToU tariff for the next 24 hours and waits for the residential agents' response. Subsequently, the suggested policy is improved upon receiving the feedback in terms of aggregated energy consumption profile for the next following episode. The simulation starts with a conventional pricing scheme where the energy cost is $\xi = 10¢/kWh$ and the DRA offers a discount price tariff every day, as an incentive for the residential agents. Fig. 3 presents a schematic diagram of the interaction between RL agents that has been developed

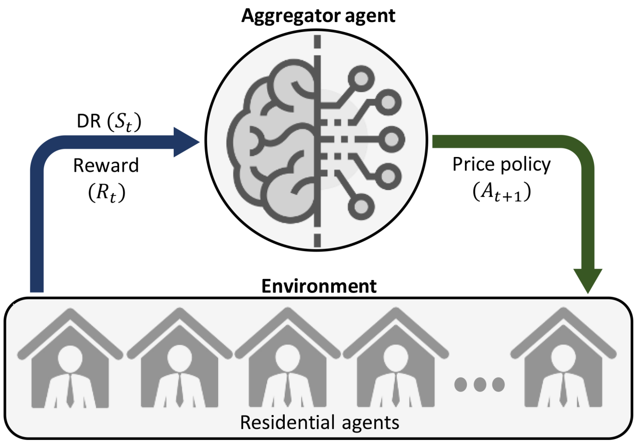


FIGURE 3. The RL environment developed using Gym toolkit.

by Gym. The results of the simulation process are discussed within the following subsections.

B. OFFLINE TRAINING RESULTS

The pre-training phase, explained above, is processed in an offline manner. In the first step of the offline training, the aggregator agent intends to determine a near-to-optimal price policy for the initial day. The learning phase starts by selecting poor actions due to the lack of knowledge. However, the reward increases at each iteration as the agent gradually gains experience. This primary aim is accomplished after 1000 episodes as demonstrated in Fig. 4. The RL convergence under all scenarios, illustrated in this figure, demonstrates that the proposed RL-based DRA can deal with the lack of information and define the near-to-optimal ToU price policies by utilizing only the DR. In fact, it is capable to deal with uncertainties related to the absence of households' internal information, for example, comfort preferences and energy flexibility potentials. Besides, it can be observed that the choice of τ is important for an optimal application of the designed structure since it affects the convergence point. Its higher values can notably restrict the changes in energy consumption and avoid improving the load factor. On the other hand, its lower amounts can bring about opportunistic residential agents and challenge sensible convergence of the results.

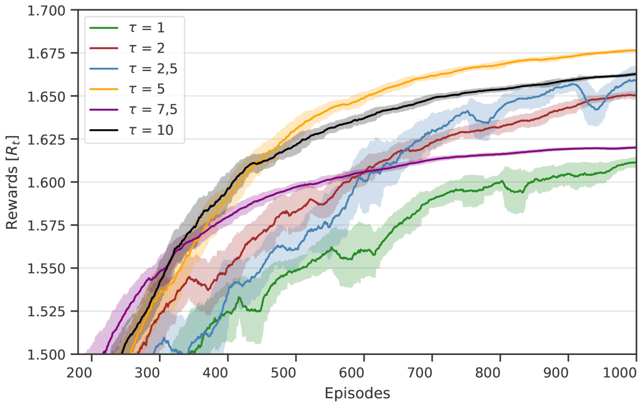


FIGURE 4. Rewards achieved by the aggregator agent within 1000 episodes of the offline training phase for different amounts of τ.

Afterward, the aggregator agent is used to generate the DB-ToU tariff for the same (initial) day, as shown in Fig. 5. As it can be seen, the generated policy, presented by the dark-red line, is able to mitigate energy consumption peaks and improve load factor. This implies the aggregator's capability to learn the DR of the residential agents. In addition, it can be observed that the recommended policy can successfully avoid any erroneous penalty to users since it maintains the energy price under the initial flat tariff while minimizing the reduction in the aggregator's income. In addition, Fig. 6 shows the difference between indoor and set-point temperatures of the house during 24 hours under the DB-ToU tariff. It can be observed that the thermal comfort of residential agents is not highly affected although the aggregated energy consumption profile is significantly altered. Particularly, the generated tariff can efficiently manage the aggregated demand by exploiting energy flexibility potentials, characterized by customer thermal comfort needs. Such management results in higher deviations from set-point temperature (notable difference) during periods with lower comfort levels while maintaining customer preferences over time with higher comfort rates (close to zero difference).

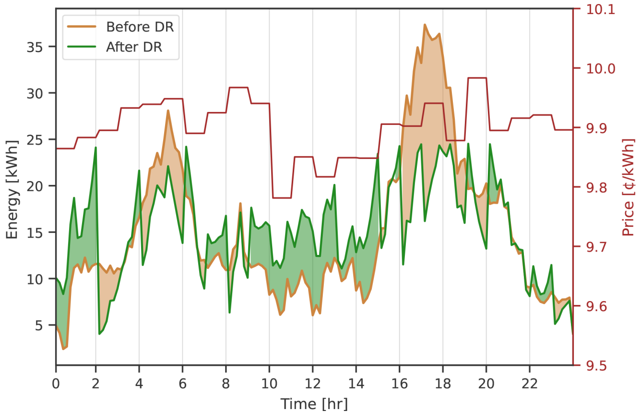


FIGURE 5. Aggregated energy consumption under the ToU tariff resulted from the offline learning process.

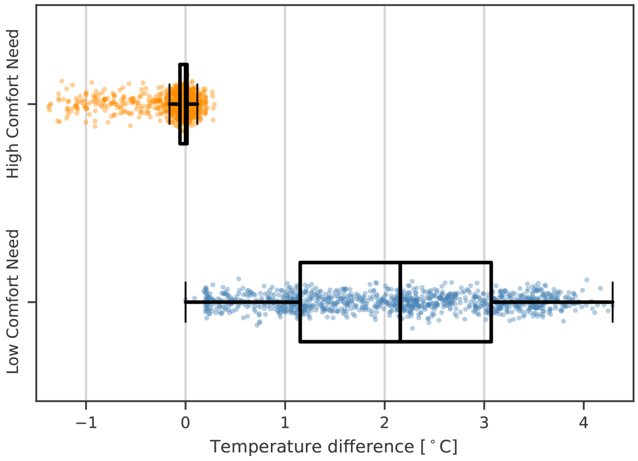


FIGURE 6. Indoor temperature deviation from set-point (temperature difference) under the DB-ToU tariff according to thermal comfort preferences of the residential agents.

Moreover, a comparative study is conducted to evaluate the performance of the proposed PPO approach in the offline training phase. For this purpose, the Deep Deterministic Policy Gradient (DDPG) and the Advantage Actor-Critic (A2C) as popular RL methods as well as a coordination technique are considered. The comparison results with the RL algorithms are presented in Fig. 7. It can be seen that PPO outperforms other techniques by higher and faster convergence. The inadequacy of DDPG can be attributed to the complexity of managing the DB-ToU tariffs across 24 hours. On the other side, A2C that starts with an inferior performance is able to

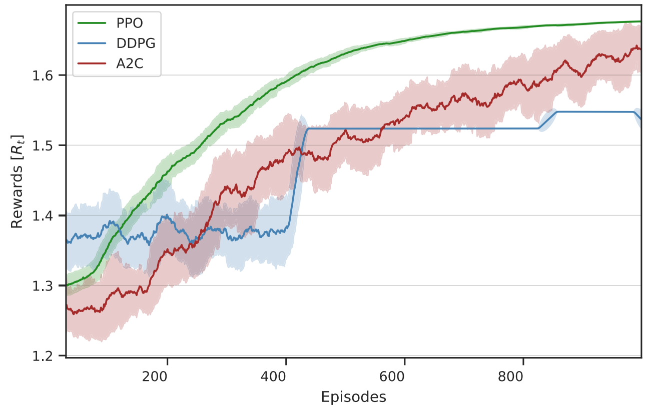
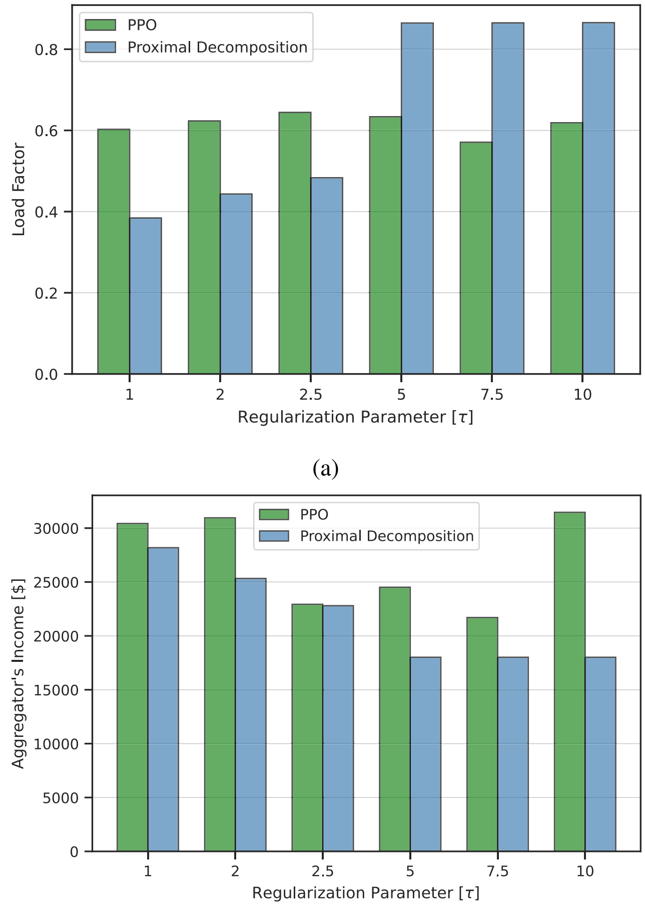


FIGURE 7. Performance comparison between different RL algorithms.



(a)

(b)

FIGURE 8. Comparison results between the proposed PPO and a coordination-based method through the offline training phase based on the load factor rate and electricity bill for different values of τ.

converge to a solution better than DDPG. Moreover, Fig. 8 illustrates the PPO outcomes in terms of the load factor rate and electricity bill for different amounts of τ compared to a coordination method, discussed in [11]. This scheme, used to coordinate residential houses, is based on non-cooperative game theory and a proximal decomposition algorithm. The proximal decomposition approach utilizes a billing mechanism proportional to aggregated demand in order to define the price policy regarding the coordination task. It can be seen in Fig. 7 (a) that the proposed PPO performs better only for the lower values of τ considering the load factor results. Nevertheless, it realizes a lower reduction in the aggregator's income for all values of τ as shown in Fig. 7 (b). The DRA can achieve such a low reduction, although the near-to-optimal tariff is based on discounts. On the other hand, a larger income reduction based on the proximal decomposition method evidence that the monetary sacrifice in a DR program can be high if it is not controlled.

C. ONLINE PERFORMANCE

Subsequently, the aggregator agent, prepared by the offline learning procedure, is deployed for consecutive days in order to evaluate its online performance. Different external temperature profiles, selected randomly from the database, are used for the evaluation. The performance comparison between scenarios with and without the aggregator agent pre-training is presented in Fig. 9. It can be recognized that the proposed pre-training system, applied to a single day, can significantly improve the efficiency of the PPO algorithm. It has reduced the convergence period from more than 1000 to a couple of days. This remarkable improvement is achieved by realizing a trade-off between choosing exploratory actions and exploiting optimal ones, defined by the aggregator agent during offline training. This strategy allows to deal with the convergence-time problem of the RL mechanisms and facilitates the future implementation of the proposed DR program.

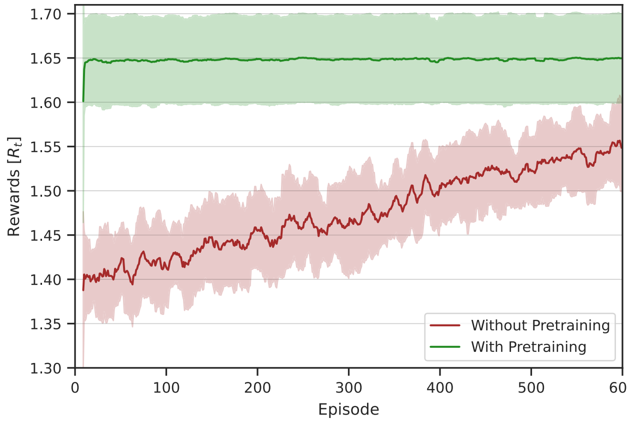


FIGURE 9. The proposed PPO performance with and without utilizing the pre-training process.

IV. CONCLUSION

This work has developed a data-driven based DRA for generating near-to-optimal DB-ToU tariffs. The proposed approach offers a DR service where the aggregator agent determines price policies based on discounts, captured by minimal information exchange with end-user agents. The suggested design reduces infrastructural needs for communication and maintains customer agents' privacy within reliable interactions. The method has recommended an RL algorithm for constructing a promising DR system. Additionally, it has carried out an offline training phase that notably improves the performance of the aggregator agent in realizing a trade-off between load factor and total revenue as two contrary objectives. As a notable achievement, this practice has avoided the time-consuming convergence of the RL and, in turn, enabled an online implementation. A comparative study with two common RL techniques and a proximal decomposition-based coordination scheme demonstrates the efficiency of the proposed DR system. Particularly, the comparison manifests the superior performance of the recommended structure through high and fast convergence rates. Future work focuses on DR studies about heterogeneous residential agents with regard to real-world applications.

ACKNOWLEDGMENT

The authors would like to thank the Laboratoire des Technologies de l'Énergie d'Hydro-Québec, the Natural Science and Engineering Research Council of Canada, and the Foundation of Université du Québec à Trois-Rivières.

REFERENCES

[1] L. Niamir, T. Filatova, A. Voinov, and H. Bressers, ''Transition to low-carbon economy: Assessing cumulative impacts of individual behavioral changes,'' *Energy Policy*, vol. 118, pp. 325–345, Jul. 2018, doi: 10.1016/j.enpol.2018.03.045.

[2] R. Lu and S. H. Hong, ''Incentive-based demand response for smart grid with reinforcement learning and deep neural network,'' *Appl. Energy*, vol. 236, pp. 937–949, Feb. 2019, doi: 10.1016/j.apenergy.2018.12.061.

[3] R. Lu, S. H. Hong, and X. Zhang, ''A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach,'' *Appl. Energy*, vol. 220, pp. 220–230, Jun. 2018, doi: 10.1016/j.apenergy.2018.03.072.

[4] P. Yang, G. Tang, and A. Nehorai, ''A game-theoretic approach for optimal time-of-use electricity pricing,'' *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 884–892, May 2013, doi: 10.1109/TPWRS.2012.2207134.

[5] K. T. Ponds, A. Arefi, A. Sayigh, and G. Ledwich, ''Aggregator of demand response for renewable integration and customer engagement: Strengths, weaknesses, opportunities, and threats,'' *Energies*, vol. 11, no. 9, p. 2391, Sep. 2018, doi: 10.3390/en11092391.

[6] V. Venizelou, N. Philippou, M. Hadjipanayi, G. Makrides, V. Efthymiou, and G. E. Georghiou, ''Development of a novel time-of-use tariff algorithm for residential prosumer price-based demand side management,'' *Energy*, vol. 142, pp. 633–646, Jan. 2018, doi: 10.1016/j.energy.2017.10.068.

[7] A. Khalid, N. Javaid, M. Ilahi, T. Saba, and A. Rehman, and A. Mateen, ''Enhanced time-of-use electricity price rate using game theory,'' *Electronics*, vol. 8, p. 48, Jan. 2019, doi: 10.3390/electronics8010048.

[8] B. Celik, R. Roche, D. Bouquain, and A. en Miraoui, ''Coordinated home energy management in community microgrids with energy sharing among smart Homes,'' in *ELECTRIMACS*. Toulouse, France: HAL Science Ouverte, 2017. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01624464

[9] B. Celik, R. Roche, D. Bouquain, and A. Miraoui, ''Coordinated energy management using agents in neighborhood areas with RES and storage,'' in *Proc. IEEE Int. Energy Conf. (ENERGYCON)*, Apr. 2016, pp. 1–6, doi: 10.1109/ENERGYCON.2016.7514081.

[10] B. Celik, R. Roche, D. Bouquain, and A. Miraoui, ''Decentralized neighborhood energy management with coordinated smart home energy sharing,'' *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6387–6397, Nov. 2018, doi: 10.1109/TSG.2017.2710358.

[11] H. K. Nguyen, J. B. Song, and Z. Han, ''Distributed demand side management with energy storage in smart grid,'' *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 12, pp. 3346–3357, Dec. 2015, doi: 10.1109/tpds.2014.2372781.

[12] S. Burger, J. P. Chaves-Ávila, C. Batlle, and I. J. Pérez-Arriaga, ''A review of the value of aggregators in electricity systems,'' *Renew. Sustain. Energy Rev.*, vol. 77, pp. 395–405, Sep. 2017, doi: 10.1016/j.rser.2017.04.014.

[13] H. Qiu, W. Gu, L. Wang, G. Pan, Y. Xu, and Z. Wu, ''Trilayer Stackelberg game approach for robustly power management in community grids,'' *IEEE Trans. Ind. Informat.*, vol. 17, no. 6, pp. 4073–4083, Jun. 2021, doi: 10.1109/TII.2020.3015733.

[14] K. Amasyali, Y. Chen, B. Telsang, M. Olama, and S. M. Djouadi, ''Hierarchical model-free transactional control of building loads to support grid services,'' *IEEE Access*, vol. 8, pp. 219367–219377, 2020, doi: 10.1109/ACCESS.2020.3041180.

[15] C. Feng, Z. Li, M. Shahidehpour, F. Wen, and Q. Li, ''Stackelberg game based transactive pricing for optimal demand response in power distribution systems,'' *Int. J. Electr. Power Energy Syst.*, vol. 118, Jun. 2020, Art. no. 105764, doi: 10.1016/j.ijepes.2019.105764.

[16] H. Taherian, M. R. Aghaebrahimi, L. Baringo, and S. R. Goldani, ''Optimal dynamic pricing for an electricity retailer in the price-responsive environment of smart grid,'' *Int. J. Electr. Power Energy Syst.*, vol. 130, Sep. 2021, Art. no. 107004, doi: 10.1016/j.ijepes.2021.107004.

[17] K. Aurangzeb, S. Aslam, S. M. Mohsin, and M. Alhussein, ''A fair pricing mechanism in smart grids for low energy consumption users,'' *IEEE Access*, vol. 9, pp. 22035–22044, 2021, doi: 10.1109/ACCESS.2021.3056035.

[18] G. Tsaousoglou, N. Efthymiopoulos, P. Makris, and E. Varvarigos, ''Personalized real time pricing for efficient and fair demand response in energy cooperatives and highly competitive flexibility markets,'' *J. Modern Power Syst. Clean Energy*, vol. 7, no. 1, pp. 151–162, Oct. 2018, doi: 10.1007/s40565-018-0426-0.

[19] H. T. Javed, M. O. Beg, H. Mujtaba, H. Majeed, and M. Asim, ''Fairness in real-time energy pricing for smart grid using unsupervised learning,'' *Comput. J.*, vol. 62, no. 3, pp. 414–429, Jul. 2018, doi: 10.1093/comjnl/bxy071.

[20] S. Datchanamoorthy, S. Kumar, Y. Ozturk, and G. Lee, ''Optimal time-of-use pricing for residential load control,'' in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Oct. 2011, pp. 375–380, doi: 10.1109/SmartGridComm.2011.6102350.

[21] N. Zhao, B. Wang, and M. Wang, ''A model for multi-energy demand response with its application in optimal TOU price,'' *Energies*, vol. 12, no. 6, p. 994, Mar. 2019, doi: 10.3390/en12060994.

[22] E. Dehnavi and H. Abdi, ''Optimal pricing in time of use demand response by integrating with dynamic economic dispatch problem,'' *Energy*, vol. 109, pp. 1086–1094, Aug. 2016, doi: 10.1016/j.energy.2016.05.024.

[23] L. D. Collins and R. H. Middleton, ''Distributed demand peak reduction with non-cooperative players and minimal communication,'' *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 153–162, Jan. 2019, doi: 10.1109/TSG.2017.2734113.

[24] B. Claessens, P. Vrancx, and F. Ruelens, ''Convolutional neural networks for automatic state-time feature extraction in reinforcement learning applied to residential load control,'' *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3259–3269, Jul. 2018, doi: 10.1109/TSG.2016.2629450.

[25] H.-M. Chung, S. Maharjan, Y. Zhang, and F. Eliassen, ''Distributed deep reinforcement learning for intelligent load scheduling in residential smart grids,'' *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2752–2763, Apr. 2021, doi: 10.1109/TII.2020.3007167.

[26] Z. Yan and Y. Xu, ''Data-driven load frequency control for stochastic power systems: A deep reinforcement learning method with continuous action search,'' *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1653–1656, Mar. 2019, doi: 10.1109/TPWRS.2018.2881359.

[27] T. Lu, X. Chen, M. B. McElroy, C. P. Nielsen, Q. Wu, and Q. Ai, ''A reinforcement learning-based decision system for electricity pricing plan selection by smart grid end users,'' *IEEE Trans. Smart Grid*, vol. 12, no. 3, pp. 2176–2187, May 2021, doi: 10.1109/TSG.2020.3027728.

[28] D. Cao, W. Hu, J. Zhao, G. Zhang, B. Zhang, Z. Liu, Z. Chen, and F. Blaabjerg, ''Reinforcement learning and its applications in modern power and energy systems: A review,'' *J. Modern Power Syst. Clean Energy*, vol. 8, no. 6, pp. 1029–1042, Nov. 2020, doi: 10.35833/MPCE.2020.000552.

[29] B.-G. Kim, Y. Zhang, M. van der Schaar, and J.-W. Lee, ''Dynamic pricing for smart grid with reinforcement learning,'' in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2014, pp. 640–645, doi: 10.1109/INFCOMW.2014.6849306.

[30] Y. Du and F. Li, ''Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning,'' *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1066–1076, Mar. 2020, doi: 10.1109/TSG.2019.2930299.

[31] L. Gkatzikis, I. Koutsopoulos, and T. Salonidis, ''The role of aggregators in smart grid demand response markets,'' *IEEE J. Sel. Areas Commun.*, vol. 31, no. 7, pp. 1247–1257, Jul. 2013, doi: 10.1109/JSAC.2013.130708.

[32] D. Azuatalam, W.-L. Lee, F. de Nijs, and A. Liebman, "Reinforcement learning for whole-building HVAC control and demand response," *Energy AI*, vol. 2, Nov. 2020, Art. no. 100020, doi: 10.1016/j.egyai.2020.100020.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[34] D. Chen, K. Chen, Z. Li, T. Chu, R. Yao, F. Qiu, and K. Lin, "PowerNet: Multi-agent deep reinforcement learning for scalable powergrid control," *IEEE Trans. Power Syst.*, vol. 37, no. 2, pp. 1007–1017, Mar. 2022, doi: 10.1109/TPWRS.2021.3100898.

[35] N. Bougie and R. Ichise, "Towards interpretable reinforcement learning with state abstraction driven by external knowledge," *IEICE Trans. Inf. Syst.*, vol. 103, no. 10, pp. 2143–2153, Oct. 2020, doi: 10.1587/transinf.2019edp7170.

[36] L. A. Hurtado, E. Mocanu, P. H. Nguyen, M. Gibescu, and R. I. G. Kamphuis, "Enabling cooperative behavior for building demand response based on extended joint action learning," *IEEE Trans. Ind. Informat.*, vol. 14, no. 1, pp. 127–136, Jan. 2018, doi: 10.1109/TII.2017.2753408.

[37] N. Bougie, L. K. Cheng, and R. Ichise, "Combining deep reinforcement learning with prior knowledge and reasoning," *ACM SIGAPP Appl. Comput. Rev.*, vol. 18, no. 2, pp. 33–45, Jul. 2018, doi: 10.1145/3243064.3243067.

[38] A. Asadulaev, I. Kuznetsov, G. Stein, and A. Filchenkov, "Exploring and exploiting conditioning of reinforcement learning agents," *IEEE Access*, vol. 8, pp. 211951–211960, 2020, doi: 10.1109/ACCESS.2020.3037276.

[39] F. Ruelens, B. J. Claessens, S. Quaiyum, B. De Schutter, R. Babuška, and R. Belmans, "Reinforcement learning applied to an electric water heater: From theory to practice," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3792–3800, Jul. 2018, doi: 10.1109/TSG.2016.2640184.

[40] J. Jeong and H. Kim, "DeepComp: Deep reinforcement learning based renewable energy error compensable forecasting," *Appl. Energy*, vol. 294, Jul. 2021, Art. no. 116970, doi: 10.1016/j.apenergy.2021.116970.

[41] J.-T. Bernard, D. Bolduc, and N.-D. Yameogo, "A pseudo-panel data model of household electricity demand," *Resource Energy Econ.*, vol. 33, no. 1, pp. 315–325, Jan. 2011, doi: 10.1016/j.reseneeco.2010.07.002.

[42] E. F. Camacho and C Bordons, *Model Predictive Control*. London, U.K.: Springer, 2004.

[43] N. Li, L. Chen, and S. H. Low, "Optimal demand response based on utility maximization in power networks," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Jul. 2011, pp. 1–8, doi: 10.1109/PES.2011.6039082.

[44] D. Toquica, K. Agbossou, N. Henao, R. Malhamé, S. Kelouwani, and F. Amara, "Prevision and planning for residential agents in a transactive energy environment," *Smart Energy*, vol. 2, May 2021, Art. no. 100019, doi: 10.1016/j.segy.2021.100019.

[45] N. F. Henao, M. Fournier, and S. en Kelouwani, "Characterizing smart thermostats operation in residential zoned heating systems and its impact on energy saving metrics," eSim, Montreal, QC, Canada, Tech. Rep., 2018. [Online]. Available: http://www.ibpsa.org/proceedings/eSimPapers/2018/1-1-A-3.pdf

[46] A. Fraija, K. Agbossou, N. Henao, and S. Kelouwani, "Peak-to-average ratio analysis of a load aggregator for incentive-based demand response," in *Proc. IEEE 29th Int. Symp. Ind. Electron. (ISIE)*, Jun. 2020, pp. 953–958, doi: 10.1109/ISIE45063.2020.9152474.

[47] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, no. 83, pp. 1–5, Apr. 2016.

3.2 Price generator function

3.2.1 Background

With the multi-agent system established, it is clear which are the minimum requirements for the different market players to ensure convergence of the price policy generation process. Now, a dynamic pricing mechanism will be developed to enhance the exploitation of the flexibility potentials from the demand side, as a more detailed price policy will be established in the day-ahead market. The idea is to determine a price policy considering supply constraints where the DSO will define a specific objective for the DRA regarding capacity limitations, while the DRA will achieve its goals considering market price constraints in the price policy generation process.

The main objective of this phase is to derive a dynamic pricing mechanism to offer a capacity limitation service to the DSO. Capacity services in a pricing context are usually offered through bidding mechanisms, leading to high computational costs and an over-reliance on customer information. In addition, there exists a lack of consideration of price limits in the literature, which could significantly impact the optimization processes of the existing approaches. This will make transitioning to a smart grid context difficult due to current regulatory and tariff structures, particularly for residential customers.

For this purpose, a dynamic price generator function is proposed, considering supply and market constraints in a game theoretic scenario implementing a coordination loop. With this function, the DRA will be able to maintain the DSO's capacity needs. In response, the DSO will pay an incentive to the DRA for maintaining capacity limits, and the DRA will try to maximize its welfare, considering both the profit from selling the electricity to the customers and the DSO's incentive. Furthermore, implementing this price generator function results in a parameterization of the price policy, reducing computational complexity. Then, the DRA will utilize an RL technique to set the parameters of this function along the coordination loop, considering not only the DSO's need but also the possible deviation of the customers' consumption profile from their stipulated consumption plans to avoid the overruns of the capacity limit.

3.2.2 Methodology

The DSO interacts with a DRA agent in order to manage the load flexibility of a group of residences. The DSO communicates the target capacity limit and offers an incentive to the DRA based on a quadratic power generation cost function. This incentive is determined based on the reduction of the power generation cost due to the peak shaving reduction. Then, the DRA utilizes the proposed price generator function in an iterative coordination loop at the beginning of the day. The DRA initially communicates a constant price profile and waits for residential agents' response. With this aggregated profile, the DRA calculates the next price policy using the price generation function until the agreement is reached.

The combination of this price generator function in the developed multi-agent system, with the coordination loop, creates a tendency that makes the maximum consumption peak lie within a neighborhood centered at the capacity limit established by the DSO with a radius that depends on the users' elasticity level. With this behavior, the following steps were followed:

1. The evaluation of the proposed price generator function regarding peak-shaving reduction.
2. A performance comparison of the proposed price generator function with an approximated piece-wise linear function in terms of overruns of the capacity limit

and flexibility exploitation.

3. An RL technique comparison between the selected PPO mechanism and the popular A2C method.
4. The evaluation of the RL-based DRA to deal with customers' consumption deviations while maximizing its welfare.

The reinforcement learning environment is composed of a set of twenty residential houses. The residential agents are considered as in the first approach using the same strategies to build the thermal models and construct the overall power profile for each house.

3.2.3 Outcomes

This work provides a price generator function to parameterize the dynamic pricing policy generation process. This function demonstrates higher and more controller exploitation of the demand side flexibility potential, enabling the offering of a capacity service for the DSO. Furthermore, it considers existing market regulations in the generation of the dynamic pricing rates, ensuring the implementation of this mechanism in realistic energy market contexts.

Simulations are carried out to evaluate the performance of the proposed RL-based strategy. This mechanism was able to exploit residential agents' flexibility to maximize DRA's profit while adjusting the parameters of the price generator function. The implementation of the RL method demonstrates that the proposed DRA can deal with agents' deviations from their consumption plans, while at the same time, the utilization of the price generator function was improved, as the DRA's profits were increased by more than 30%. Regarding RL mechanism selection, the adopted PPO method converged to a solution that provides higher rewards for the DRA than the well-known A2C method.

Smart Energy 14 (2024) 100139

Contents lists available at ScienceDirect

Smart Energy

journal homepage: www.journals.elsevier.com/smart-energy

Deep reinforcement learning based dynamic pricing for demand response considering market and supply constraints

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

[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
Market constraints
Capacity limitation
Reinforcement learning

ABSTRACT

This paper presents a Reinforcement Learning (RL) approach to a price-based Demand Response (DR) program. The proposed framework manages a dynamic pricing scheme considering constraints from the supply and market side. Under these constraints, a DR Aggregator (DRA) is designed that takes advantage of a price generator function to establish a desirable power capacity through a coordination loop. Subsequently, a multi-agent system is suggested to exploit the flexibility potential of the residential sector to modify consumption patterns utilizing the relevant price policy. Specifically, electrical space heaters as flexible loads are employed to cope with the created policy by reducing energy costs while maintaining customers' comfort preferences. In addition, the developed mechanism is capable of dealing with deviations from the optimal consumption plan determined by residential agents at the beginning of the day. The DRA applies an RL method to handle such occurrences while maximizing its profits by adjusting the parameters of the price generator function at each iteration. A comparative study is also carried out for the proposed price-based DR and the RL-based DRA. The results demonstrate the efficiency of the suggested DR program to offer a power capacity that can maximize the profit of the aggregator and meet the needs of residential agents while preserving the constraints of the system.

1. Introduction

Demand-side management plays a key role in optimizing end-users' demand in smart grids. This idea facilitates power system operation through different services, including the liberalization of electricity markets, real-time balance of demand and supply, the improvement of load control strategies, the reduction of energy consumption, and the integration of decentralized energy resources [1]. Accordingly, it assists the smart grid with the self-optimization concept (distributed optimization) that promotes more continuous and sophisticated demand-side participation. Particularly, Demand Response (DR) programs, as an important facet of demand-side management, enable the management of various controllable and programmable loads in the residential sector, such as thermostatic devices, plug-in electric vehicles, and smart appliances [2]. This energy flexibility program leads to the realization of smart distribution grids where residential customers participate in grid operation as active players [3].

The DR programs have been developed to mitigate peak load by changing consumption patterns in response to price or incentive signals [4,5]. Monetary incentives influence clients to modify their load profiles without significantly compromising their comfort preferences [6]. From a realistic standpoint, peak demand management is crucial to power system reliability regarding the designed capacity of the grid. From a financial perspective, such a service is pivotal to electricity generators that must operate with higher costs during peak periods to manage the additional usage [7]. Therefore, the reduction of peak load through implementing DR programs is a key strategy that offers benefits for both the demand and supply sides.

An effective DR program can be realized through capturing demand flexibility at its full potential. Accordingly, the DR Aggregator (DRA) has emerged as a commercial entity to explore such an opportunity by negotiating agreements between consumers and market [8]. This mediator recruits customers and directly contacts clients using information

* Corresponding author.
E-mail address: alejandro.jose.fraija.ochoa@uqtr.ca (A. Fraija).

https://doi.org/10.1016/j.segy.2024.100139
Received 5 September 2023; Received in revised form 21 March 2024; Accepted 22 March 2024
Available online 27 March 2024

Nomenclature

Acronyms

DR	Demand Response
DRA	Demand Response Aggregator
DSO	Distribution System Operator
ESH	Electric Space Heating
MDP	Markov Decision Process
PAR	Peak-to-Average Ratio
PPO	Proximal Policy Optimization
RL	Reinforcement Learning

Functions

$\hat{A}_t$	Advantage at episode t
$\psi(\cdot)$	Power generation cost reduction function
$\xi(\cdot)$	DRA welfare function
$g(\cdot)$	Thermal model
R_t	Reward function at episode t
$U(u_k^i)$	Thermal comfort function

Indices

i	House index
k	Time-step index
t	Iteration index

Parameters

α	Rate of price change
π_{max}	Upper price limit
π_{min}	Lower price limit
M	Capacity limit

Variables

δ_k^i	Thermal discomfort factor of i^{th} house
η	Capacity limit reduction
$\hat{u}_k^i$	Actual energy consumption of i^{th} house at time-step k
μ_t^h	Normalized aggregated consumption
a_t	Action at episode t
s_t	State at episode t
u_k^i	Energy consumption reported of i^{th} house at time-step k
x_k^i	Indoor temperature of i^{th} house at time-step k
x_k^{out}	Outdoor temperature at time-step k
x_{comf}^i	Set-point temperature profile of i^{th} house
y_k	Aggregated energy consumption time-step k

and communication technologies [9]. As a result, it collects load flexibility and offers it as a service to the Distribution System Operator (DSO). Congestion management, power quality improvement, and grid capacity expansion are critical exercises performed by the DSO based on this flexibility [10,3].

Specifically in the residential sector, an important source of flexibility is the thermal loads [11]. In countries with harsh winters, residential thermal loads are among the major energy-expensive appliances. For instance, in Quebec, Electric Space Heating (ESH) systems account for about 60% of household energy consumption [12]. These appliances can cause a significant increase in power demand during peak load and, at the same time, represent a critical factor in the user's electricity bill. Because of this, smart programmable thermostats are widely employed to manage the problems, from the user's point of view, of reducing their electricity bills. Alternatively, these controllable devices release the opportunity to capture the flexibility potentials of these loads, which can be capitalized by the DRA, enabling new possibilities for both the demand side and the DRA that can be exploited through the implementation of DR programs [13].

One of the key elements in the correct implementation of DR programs in the residential sector, is the optimal generation of price-based policies [14]. The main goal of these mechanisms is to exploit the flexibility potential from the demand side to deal with the problem of consumption peaks. However, there exist some challenges for the DRA in implementing these mechanisms at the residential level, starting with significant privacy concerns [15], resulting in affecting the optimality of DR policies due to the uncertainty that comes from the lack of information provided by the user, like users' thermal comfort preferences [16]. Moreover, if the problem is analyzed from the grid perspective, performing this exercise without considering the needs of the network can generate imbalances in the system, as shown in [17]. In addition, existing market regulations establish limits for the sale of energy, which makes most of the studies that do not consider restrictions on price generation unsuitable for retailers such as DRAs [18]. This is evidence of the need to continue exploring these types of scenarios to avoid a myopic generation of pricing tariffs that end up affecting the grid stability or in unprofitable strategies for the DRA.

In this regard, this research study addresses optimizing thermal energy usage among a group of residential customers considering a DRA despite supply and market constraints. It tackles this issue by introducing a price generator function that utilizes the aggregated consumption profile as the only source of information to generate price policies. Furthermore, the function takes into account the existing market regulations to establish restrictions in a dynamic pricing approach, and allows the translation of a target capacity limit into a dynamic pricing policy through a coordination process. As a result, this mechanism proves its capabilities at exploiting residential flexibility in a controlled manner, and reducing power generation costs while simultaneously increasing the profit for the DRA. To set the function parameters that optimize the generation of price-based policies through the coordination loop, a reinforcement learning (RL) mechanism is used to deal with the lack of information regarding the users' objectives. The RL mechanism is implemented for two reasons, first, it allows dealing with the complex environment with incomplete information on the DR program, and second, it will handle the users' deviations in the execution of the consumption plans to guarantee the respect of the capacity limit stipulated by the DSO.

1.1. Related works

Price-based DR programs are formulated to deal with the challenges of defining prices/rates for different time blocks in an optimal manner, especially in day-ahead markets [19]. In fact, the idea of offering fixed prices to residential customers for long periods in order to maintain the balance of the power grid as a complex real-time system can yield inefficient performances [20]. In this regard, the implementation of dynamic pricing schemes is suggested that can provide an efficient utilization of generation capacity. These strategies encourage users to change their consumption patterns without modifying generators' costly operation [21]. Nevertheless, acquiring an optimal pricing design is difficult due to inherent uncertainties in DR programs related to customers' dynamic load consumption and price-responsive behavior. For instance, the authors in [22,23] have addressed this situation by developing optimal dynamic pricing mechanisms that allow a trade-off between consumers and the utility. Their method has roots in the two most popular practices in price-based DR programs. The first performs optimization problems that rely on an extensive exchange of specific information [22,24,25]. Subsequently, in many cases, they can affect the privacy and participation interests of customers. The second implements iterative processes commonly based on game theoretical frameworks [23,26,27]. The over-reliance of these procedures on users can give them opportunities to game the system. In response to these issues, in [28,29], the authors have proposed non-cooperative approaches to reduce the peak of aggregated energy consumption profile. A similar strategy that shares the power consumption cost between users has been suggested by the authors in [29]. However, these solutions suffer from the lack of constraints on price generators that can result in either unwanted penalties against users or barriers to implementing constrained markets.

On the other hand, the emergence of DRA in the implementation of DR services has allowed different approaches to be explored. The interactions between these entities and households have also enabled the development of markets with capacity constraints. As an example, the authors in [30] took advantage of this interaction to impose capacity constraints, in which they propose a strategy for constructing a bidding curve for capacity increments. In this regard, in [31] a market-clearing mechanism was developed for offering a capacity limitation service. This work investigates at what costs aggregators can offer capacity constraints, and how these can reduce the DSO's network operating cost. These bidding mechanisms have a good response in capacity-constrained flexibility markets. However, the need for intrusive approaches to the construction of aggregators' bidding models can be a disadvantage in their implementation. Moreover, the additional workload for DSOs to submit or clear bids in these markets remains a major obstacle to their implementation. In this regard, authors in [32] proposed a mathematical framework for a dynamic pricing mechanism in an energy community to enable the provision of capacity limitation services to the DSO. They highlight the importance of extending the portfolio of local flexibility resources to thermostatically controlled loads. However, no price limits have been taken into account, and the suggestion of a bi-level optimization may result in privacy issues from the demand side.

Recently, researchers have focused their efforts on utilizing Reinforcement Learning (RL) methods in order to solve the existing issues. Particularly, an RL agent can handle system uncertainties without any prior knowledge [33]. The approach of the authors in [34,35] relies on employing the RL technique for an optimization problem with a combined objective function to meet the desires of both consumers and the aggregator in a real-time context. However, such a manner of formulating users' preferences raises privacy issues since it requires access to their dissatisfaction information during the price policy generation process. In a previous study, the authors have addressed this obstacle by developing a learning procedure only based on the aggregated load to define RL actions, and thus, alleviated privacy concerns [36,37]. The related research also considered price constraints determined by the market to improve either the Peak-to-Average ratio (PAR) or the Load Factor. Although there are significant achievements in terms of flattening the energy consumption curve by means of RL techniques, there is no clear link between peak reduction and system balance. This highlights the need to explore a different approach that allows for utilizing end-users flexibility in a controlled way based on the maximum consumption expected by the DSO. Such consideration brings about an optimal means to facilitate maintaining the power grid's reliability.

1.2. Motivation and contribution

The main objective of this paper is to derive a dynamic pricing mechanism to provide a capacity limitation service considering the established energy market regulations. For brevity of the presentation, Table 1 compares the differences between the existing methods and the proposed model, demonstrating the lack of consideration of price limits in the literature, which could significantly impact the optimization processes. In addition, capacity services in a pricing context are usually offered through bidding mechanisms, which leads to high computational costs and an over-reliance on the information provided by customers. These points are a further barrier to DR program implementations [18] related to current regulatory and tariff structures, particularly for residential customers. Moreover, one of the remaining fundamental issues is pricing in a demand response scenario of the power market by respecting both the capacity and operational costs of responding.

To overcome the aforementioned issue and develop a dynamic pricing mechanism, we introduce a price generator function for the DRA by considering power capacity and market constraints. Each residential user independently determines its best response strategy to minimize energy costs and maximize profit. The proposed DRA uses the price generator function in a game theoretic scenario to coordinate customer responses. The proposed method takes advantage of RL techniques to estimate the price generator function parameters and a proximal decomposition algorithm as a regularizer on the customers' side. The regularization allows us to ensure the convergence of the proposed multi-agent system. Accordingly, this work contributes,

1. A price-based DR program centred on proposing a price-generating function for the DRA agents that considers the market price restrictions. This work identifies a sigmoid function that, combined with the regularization of users' DR based on proximal decomposition in a coordination loop, allows the reduction of local peaks according to the stipulated capacity limits.
2. An RL method to determine the parameters of the price generator function during the coordination loop. These parameters assist in maximizing the DRA's profit while respecting DSO's service needs. The PPO algorithm is used to overcome the lack of user information in the process of optimizing pricing policies.
3. An RL-based DRA agent that considers the deviations from consumers from their stipulated consumption plans. This agent can characterize users' variations to avoid significant impacts on the power constraints of the system while improving the DRA's profit. The data-driven mechanism makes it possible to characterize the uncertainty of user deviations during the execution of consumption plans.

The rest of the paper is organized as follows: Section 2 presents the methodology for the developed framework. Section 3 covers the validation setup. The results are discussed in Section 4, followed by the conclusion in Section 5.

2. DR mechanism and problem formulation

In a residential distribution grid, operated by automated agents, DSO interacts with a DRA agent in order to manage load flexibility of a group of residences. The DRA provides monetary incentives by managing the price policy. In response, the customers change their energy consumption patterns that helps avoid network congestion and ensure the system reliability. Indeed, this constitutes a mechanism in which customers communicate their consumption plan with the DRA in response to a stipulated price profile. Although the DRA does not know consumers' preferences in this structure, it can adapt the price profile according to their propositions. In this regard, Fig. 1 illustrates the structure of the proposed price-based DR mechanism. In the designed framework, the DRA runs the day-ahead planning of a set of residential agents. It communicates to them price signals in a coordination loop and induces them to react. Through this interaction, the DRA seeks to decrease the aggregate peak demand by regulating customers' power profiles. Specifically, the DRA defines a constant price profile and waits for the users' response. Upon receiving the feedback, the DRA adapts the price profile and waits for the residential agents' new consumption plan until reaching an agreement.

2.1. Price generator function

In order to define the DRA's price profile, a price generator function is formulated considering π_{min} and π_{max} as the market's minimum and maximum price constraints accepted for the DR mechanism. This

Table 1
Comparison between the existing methods and the proposed model regarding objective functions, consideration of capacity limitation, and price constraints.

Ref	DR Mechanism	Pricing generation Method	Objective Function	Capacity Limitations	Price Constraints	Demand side strategy
[20]	Dynamic pricing	Binary genetic algorithm	Minimize the average system cost and rebound peaks	✗	✗	Load scheduling with photovoltaic renewable energy source integration
[22]	Dynamic pricing	Multi-objective optimization	Considers the benefits and costs of the opposing entities at both ends of supply and demand	✗	✗	Energy optimization and scheduling for renewable microgrid
[23]	Dynamic pricing	Multi-objective optimization	Social welfare maximization	✗	✗	Optimal scheduling of thermostatically controlled loads
[24]	Dynamic pricing	Bi-level, meta-heuristic	Profit maximization for retail electricity provider and cost minimization for customers	✗	✗	Consumption optimization of interruptible, non-interruptible, non-shiftable, and curtailable loads.
[25]	Real-time pricing	Single-objective optimization model	Minimize the electricity cost and electricity consumption dissatisfaction	✗	✗	Energy optimization for prosumers with distributed energy and energy storage devices
[26]	Demand biding	Bi-level game-theoretic model	Maximizes the social welfare of the local power exchange market and minimizes the social cost of the day-ahead wholesale market	✗	✓	Optimal control of customers' switching behaviors
[27]	Day-ahead pricing	Stackelberg game-theoretic model	Maximize aggregator's profit	✗	✗	Flexibility level based price-responsive behavior
[28]	Time-ahead pricing	Game-theoretic model	Minimizes the player's costs based on the predicted strategy of all other players	✗	✗	Optimal charging of electric vehicles
[29]	Dynamic pricing	Game-theoretic model	Minimizes the square euclidean distance between the instantaneous load demand and the average demand for the energy provider and minimizes energy payment for the users	✗	✗	Optimal appliance scheduling and control of energy storage devices
[30]	Demand biding	Stochastic optimization	Minimizes the deviation from a baseline load profile	✓	✗	Optimal control of thermostatically controlled loads and photovoltaic generators
[31]	Demand biding	Market clearing mechanisms	Minimizes overall social cost	✓	✗	Optimal energy management strategy for their distributed energy resources
[32]	Dynamic pricing	Bi-level optimization	r minimizes the total operational cost of an energy community	✓	✗	Optimal control of production facilities and/or an energy storage system for prosumers
[34]	Dynamic pricing	Reinforcement learning	Maximizes service provider profit and minimizes customers' costs	✗	✗	Energy management of critical and curtailable loads
[35]	Dynamic pricing	Reinforcement learning	Minimizes the expected discounted system cost of the service provider	✗	✗	Minimize consumers' dissatisfaction utilizing an energy disutility function
[36]	Distribution locational marginal price	Reinforcement learning	maximize the total profit of selling power	✗	✗	A data-driven deep neural network to model a multi-microgrid price responsive behavior
[37]	Time-of-Use	Reinforcement learning	Maximizes the load factor and demand response aggregator's profit	✗	✓	Optimal control of electric space heating
[38]	Dynamic pricing	Three-tiered optimization	Maximize the financial savings from renewable energy	✗	✓	energy optimization and scheduling for renewable microgrids
[39]	Dynamic pricing	Stackelberg game-theoretic model	Maximize subcontracting power supply profit	✗	✓	Control capabilities of air-conditioning systems and electric vehicles for commercial buildings
Proposed work	Dynamic pricing	Reinforcement learning	Minimizes demand response aggregator profit reduction and the cost of exceeding the capacity limitations	✓	✓	Optimal control of electric space heating

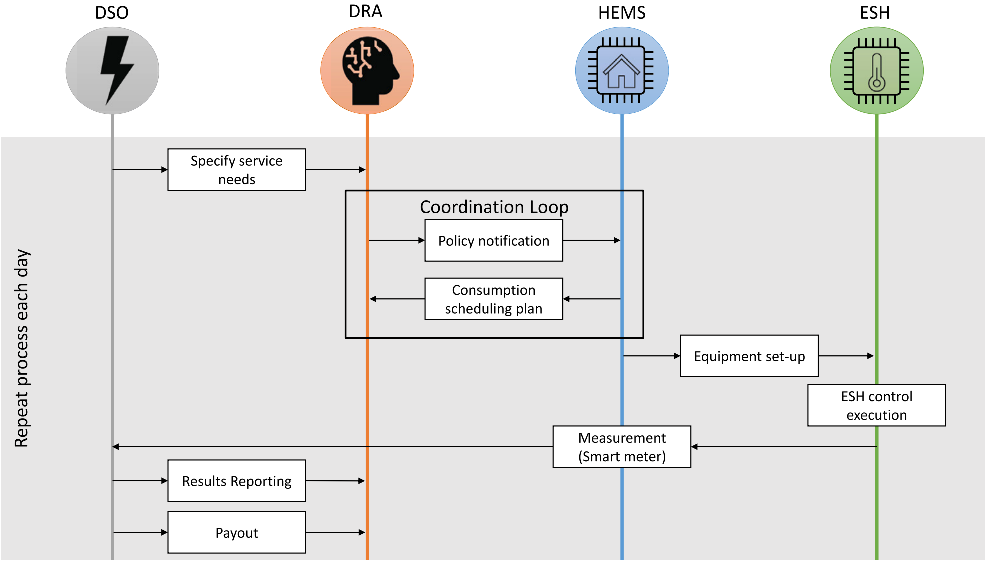


Fig. 1. Automatic price-based DR sequence.

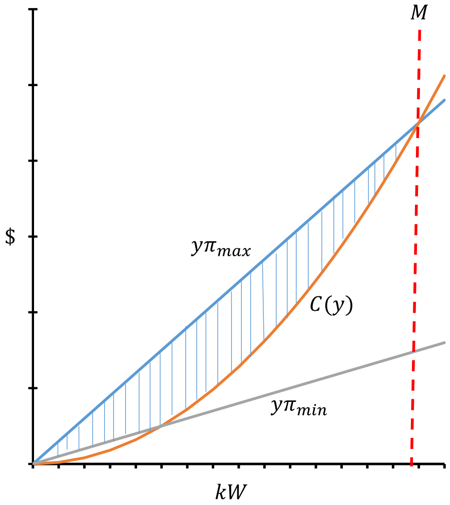


Fig. 2. Market and power constraints in terms of power generation cost function.

consideration is important as it restricts the implementation of many existing mechanisms that do not consider these price constraints in their algorithms. Then, the following price generator function allows entities like DRAs to compete in this type of market, where optimizing their profits becomes an important challenge. Moreover, the generator function considers a capacity limitation factor M established by the system. This factor is defined by the DSO based on the power generator cost function of the energy provider (see Fig. 2). This means that the DSO may define a value for M when the power grid operation is compromised. Aspects such as maintenance reduction or operating cost reduction, would determine the M value based on physical system constraints (such as maximum transformer capacity) or maximum desired node capacity (for reducing system losses), respectively. Accordingly, we propose the following price generator function,

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

where y_k represents the aggregate consumption at time stamp $k \in \{1, \dots, N\}$. This value corresponds to the sum of individual household energy consumption, i.e. $y_k = \sum_{i=1}^{H} u_k^i$, where H represents the number of houses, and u_k^i is the energy consumption of the i^{th} house at the time stamp k. Lastly, α is a positive parameter that controls the rate of price change. To properly determine this value, exploration must be conducted by the DRA agent due to the lack of existing information linked to the relationship between the users' elasticity and flexibility.

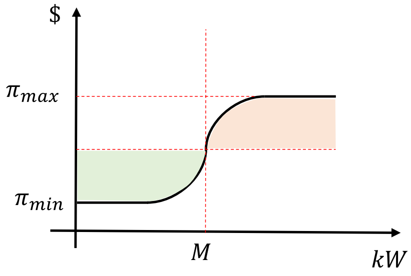


Fig. 3. Proposed price generator function.

The proposed price generator function, $\pi_k(y_k)$, has some particular properties that make it suitable for reducing aggregate load peaks of the aggregated demand profile. In fact, the developed function establishes a direct correlation between consumption and price at every time slot. This means that prices increase or decrease in the same way that aggregate consumption does.

Furthermore, the function has an inflection point at M that allows for a division into two convex regions, as shown in Fig. 3. Since users participate with their best responses, their energy payments either decrease or remain unchanged while reducing their consumption peaks. As a result, consumers try to avoid the high price region. This tendency makes $\max_k(y_k)$ lie within a neighborhood centred at M with a radius of r depending on the users' elasticity level.

2.2. DRA agent

In the described scenario, the DRA takes into account the prevailing market regulations that impose restrictions on energy unit selling prices. Additionally, the proposed approach aims to mitigate consumption peaks considering the defined objectives set by the DSO regarding capacity constraints. These limitations are accounted for in the design of the price generator function. Consequently, the DRA endeavors to maximize its profit by avoiding exceeding the stipulated capacity limit, utilizing the feedback obtained from the interaction with the residential agents.

This interaction between the set of residential agents and the DRA is modeled as a multiple-follower and one-leader Stackelberg game. In this model, the leader seeks to optimize its usefulness which depends on the profit from the electricity supply to customers and the cost of exceeding the power constraints of the system. The energy cost related to the provider can be modeled by the quadratic function $C(y_k) = ay_k^2 + by_k + c$ that has been widely used in the literature [29,40]. For this analysis, we define $a = \pi_{max}/M$ and $b = c = 0$, considering the break-even point between the cost function and the revenue produced by π_{max}. The profit depends on the price policy established by the DRA in (1), while the cost is indirectly controlled through interactions between the followers and the leader. The DSO determines the DRA reward ψ based on the cost reduction concerning the initial aggregated consumption plan, i.e.,

$$\psi = \sum_{k=1}^{N} C(y_{0,k}) - \sum_{k=1}^{N} C(y_k) \tag{2}$$

Therefore, considering $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_N\}$ as the price policy for the next interaction, the DRA benefit can be explained by the difference between its income and the cost of exceeding the power constraint,

$$\xi(\boldsymbol{\pi}) = w_1 \left(\sum_{k=1}^{N} y_k \pi_k + \psi \right) - w_2 (\max_{k=1,\ldots,N} y_k - M), \tag{3}$$

where w_1 and w_2 are weighting factors to balance these two terms. In this case, each one of these factors is defined first by the inverse of the unweighted historical average of each term to guarantee a normalized result; thereafter, these values are slightly modified to give more importance to the cost per overrun. This function (3) is difficult to optimize since it is not convex; thus, it cannot be treated by the classical gradient-based optimization methods. Moreover, the deviation from the consumption plan by the residential agents during the DR practice evidences the need for an algorithm with the ability to handle such uncertainty. Consequently, the RL method is implemented to deal with the intractability of the DRA price generation problem. RL algorithms have strong exploration capabilities that enable them to interact continuously with an unknown environment and constantly update the agents' experience towards an optimal decision [41]. Despite the drawback linked to the training time of RL algorithms, they offer the benefit of addressing nonlinearities within optimization problems, as outlined in [42]. This study illustrates how RL methods have been utilized to overcome the necessity of acquiring the dynamics of nonlinear systems for implementing optimal control strategies. The aforementioned demonstrates that employing the RL approach enables the optimization of the DRA's pricing strategy within the intended scenario.

2.2.1. An overview of the RL

RL algorithms are based on an agent interacting with an unknown environment and performing actions to extract useful information. Through these interactions, the agent attempts to maximize its reward by realizing a trade-off between exploring new actions and exploiting those that seem optimal [43]. This process starts by observing the state of the environment. The RL agent acts and receives an immediate reward and the resulting new state from the environment. This is because, during the iterative process of interactions between the RL agent and the environment, the action affects the environment causing a change in its state according to a given probability [44].

When starting the iterative process, the RL agent is unaware of the link between the action performed in a given state with the reward and the new state received as a response from the environment. In fact, the agent learns this knowledge by continuously interacting with the environment. The acquired comprehension is used by the agent to maximize not only the immediate reward but also the expectation of the future ones. It can be deduced that an RL algorithm is a trial-and-error approach that looks to optimize a decision-making process.

2.2.2. RL representation of a dynamic pricing mechanism under capacity constraints

The targeted scenario considers a multi-agent system composed of a set of residential agents and an RL-based DRA. The interactions between the residential environment and the RL agent are modeled by a Markov Decision Process. This decision-making formalism allows modeling an environment as a set of states where the states of the environment are Markovian, and actions can be performed to control the system's state for maximizing some performance criteria. This can be used to learn sequential decision-making processes by mapping states onto actions in such a way that the expected outcome will produce the intended effect. These mapping strategies are called policies in this theory. Thus, the Markov Decision Process framework enables the gradual learning of optimal policies through consecutive trials, applying different methods developed in the literature [45]. According to the aforementioned, the model is represented by a tuple $\langle S, A, P, R, \gamma \rangle$, where S and A are the sets of states and actions, respectively. P presents the state transition probability, R is a reward function, and γ stands for a discount factor [46].

The RL-based DRA defines the action $a_t \in A$ at each step according to the state $s_t \in S$. $s_t = \{\mu_{t,1}, \mu_{t,2}, \ldots, \mu_{t,N}\}$ is the normalized aggregate consumption profile, where $\mu_{t,k} = \dfrac{y_k}{\max\limits_{k \in \{1,\ldots,N\}} \{y_k\}}$. The action a_t modifies the price generator function to maximize the reward of DRA within the coordination loop. In this regard, $a_t = \{\eta, \alpha\}$ where η is a parameter established to allow the DRA to transform the price generator function for dealing with residential agents' deviations. As a result, the price generator function, $\dot{\pi}_k(.)$, utilized by the DRA and the reward function, R_t, defined for our RL set-up, can be described through (4) and (5), respectively.

$$\dot{\pi}_k(y_k, \eta, \alpha) = \pi_{min} + \frac{\pi_{max} - \pi_{min}}{1 + \exp\left(\dfrac{-y_k + M - \eta}{\alpha} \right)} \tag{4}$$

$$R_t = \xi(\dot{\boldsymbol{\pi}}) \tag{5}$$

The DRA agent determines actions that maximize its cumulative reward $G_t = \sum_j \gamma^{j-1} R_j$ as the return over a number of steps named episode. In this case, an episode is equal to the coordination loop between the DRA and residential agents.

2.2.3. Proximal policy optimization (PPO) method

The implemented RL algorithm is based on the PPO technique. This policy gradient means is used to optimize the policy $\phi_\theta(a_t, s_t)$ based on the parameter θ. The policy describes the agent's behavior as a rule to decide the action in a given state. This technique tries to stabilize the training process of the RL agent by avoiding parameter updates that can produce a high policy alteration in a single step. Additionally, it attempts to keep old and new policies as closely as possible, ensuring reward enhancement and stability during the process [47]. For this purpose, the PPO scheme maximizes an objective function, $J(\theta)$, with respect to θ, i.e.

$$J(\theta) = \hat{\mathbb{E}}_t[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)], \tag{6}$$

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 $\phi_\theta(a_t|s_t)$ / $\phi_{\theta_{old}}(a_t|s_t)$. The PPO method uses $\hat{A}_t = -V(s_t) + \gamma R_t + \cdots + \gamma^{T-t+1} R_{T-1} + \gamma^{T-t} V(s_T)$ as the estimated advantage at episode t, where T is the batch size. This advantage function measures the performance of a selected action given the current state. Finally, ϵ is the hyperparameter for clipping. This parameter avoids large deviations in the θ updating process by setting the ratio in the interval $[1 - \epsilon, 1 + \epsilon]$ [48]. The Algorithm 1 in Appendix A represents the utilized PPO technique for the targeted scenario.

2.3. Automated DR for residential agents

It is assumed that each residential agent is equipped with a home energy management system (HEMS), which enables flexible demand. In this practice, flexible load refers to heating systems controlled by smart thermostats based on end-users' comfort. The possibility to modify the thermal load provides the flexibility required for residential agents' participation in the DR program. On the other hand, fixed load refers to other household appliances operating without the same strategy.

Subsequently, the heating consumption can be computed by maximizing the individual welfare, expressed by,

$$\begin{aligned} \underset{\mathbf{u}^i=\{u_k^i\}_{k=1}^N}{\text{Maximize}} \quad & J(\mathbf{u}^i) \\ \text{subject to} \quad & x_{k+1}^i = g(x_k^i, x_k^{\text{out}}, u_{h,k}^i), \\ & x_k^i \in [x_{\min}^i, x_{\max}^i], \\ & u_k^i \in [0, u_{\max}^i], \\ & u_k^i = u_{h,k}^i + u_{a,k}^i, \end{aligned} \tag{7}$$

where the vector $\mathbf{u}^i = \{u_1^i, \cdots, u_N^i\}$ is the consumption plan of the i^{th} house. The variables x_k^i and x_k^{out} are the indoor and outdoor temperatures. $u_{h,k}^i$ stands for the heating energy consumption. The total energy consumption of the i^{th} house at the time k accounts for the aggregation of thermal and fixed loads, $u_k^i = u_{h,k}^i + u_{a,k}^i$. The thermal model of the house, $g(\cdot)$, is a discrete linear model described in [49]. The setting of this model, based on real data, is presented in Section 3. The parameters $x_{\min}^i$ and $x_{\max}^i$ are the minimum and maximum allowed internal temperatures set by the user. The objective function, $J(\mathbf{u}^i)$, is defined as,

$$J(\mathbf{u}^i) = \sum_{k=1}^{N} U(u_k^i) - \pi_k u_k^i, \tag{8}$$

where π_k represents the energy price at k and $U(u_k^i)$ is the utility function of the customer, which in this case is the thermal comfort, i.e., the goal of the user is to maintain its comfort needs while reducing its bill.

According to the literature, several methods for modeling user comfort have been proposed as presented in [50]. These models are based on ISO and ASHRAE standards to determine which are more interesting [51]. Based on this, the Fanger model is a very common analysis, that utilizes the characteristic numbers Predicted Mean Vote (PMV) and Predicted Percentage of Dissatisfied (PPD) to determine the thermal comfort of occupants, [52]. However, implementing these strategies implies using a larger number of variables, needing the utilization of more complex thermal models. This would result in a significant increase in algorithmic complexity. For this reason, without losing generality, a linear thermal model is implemented, which is computationally less demanding. The model $g(\cdot)$ for the thermal dynamics of the house, based on the indoor temperature x_k^i, the outdoor temperature x_k^{out} and the thermal consumption $u_{h,k}^i$ is defined as follows, where $\boldsymbol{\beta}^i = [\beta_1^i, \beta_2^i, \beta_3^i]$ are the state transition coefficients:

$$x_{k+1}^i = g(x_k^i, x_k^{\text{out}}, u_{h,k}^i) = \beta_1^i x_k^i + \beta_2^i x_k^{\text{out}} + \beta_3^i u_{h,k}^i. \tag{9}$$

Then, the residential agents aim to minimize their thermal comfort dissatisfaction, i.e., the difference between the desired and indoor temperature has to be minimized [53]. With this in mind, since the residential agent uses the thermal load as flexible demand, this function is determined based on thermal comfort parameters consisting of x_{comf}^i as the set-point temperature and δ_k^i as the comfort weight factor. This element represents users' ability to sacrifice comfort to reduce the bill. According to [49,54], the thermal comfort can be modeled with the following quadratic utility function,

$$U(u_k^i) = -\delta_k^i (x_{\text{comf}}^i - x_k^i)^2, \tag{10}$$

where δ_k can take two values from the set $\{0, \delta_{\max}\}$. In the case of $\delta_k = \delta_{\max}$, occupants are interested in reaching their comfortable temperature set-point. Indeed, the parameter $\delta_{\max}$ advertises the price elasticity of the heating energy. This strategy maximizes the flexibility of the residential agent without compromising its thermal comfort constraints. For instance, the agent can freely modify the internal temperature under $\delta_k = 0$ while respecting the constrain $x_k^i \in [x_{\min}^i, x_{\max}^i]$.

Since the residential agents are simultaneously solving their optimization problem in a selfish way, it is necessary to regularize their optimization problems. According to theorem 3 in [29], this regularized plan of the houses combined with the non-negative users' payments granted by the price generator function guarantees the existence of a Nash equilibrium in the proposed DR mechanism. The proximal decomposition can perform the regularization as a distributed algorithm [55]. In this regard, a regularization parameter, τ, is utilized to penalize the difference between consecutive defined consumption plans, i.e., penalize significant variations between episodes t and $t-1$ [37]. As a result, the dual optimization problem to minimize the residential agents' cost function can be defined by (11).

$$\begin{aligned} \underset{\mathbf{u}^i=\{u_k^i\}_{k=1}^N}{\text{Minimize}} \quad & \sum_{k=1}^{N} \delta_k^i (x_{\text{comf}}^i - x_k^i)^2 + \pi_k u_k^i + \tau (u_{t,k}^i - u_{t-1,k}^i)^2 \\ \text{subject to} \quad & x_{k+1}^i = g(x_k^i, x_k^{\text{out}}, u_{h,k}^i), \\ & x_k^i \in [x_{\min}^i, x_{\max}^i], \\ & u_k^i \in [0, u_{\max}^i], \\ & u_k^i = u_{h,k}^i + u_{a,k}^i. \end{aligned} \tag{11}$$

Although all customers intend to report and consume the optimal demand, which minimizes their costs, deviations can appear during run time. Such deviations indicate that users consumed d_k times their reported plan, i.e., $\hat{u}_k = d_k u_k$ at each time stamp [56]. In order to model the occurrence of such deviations, d_k can be expressed as a random variable that follows a Log-normal distribution with parameters $\mu = e$, and $\sigma = 0.05$.

3. Validation setup

In this section, the proposed DR mechanism is validated through numerical analyses. The experimental data used for constructing the thermal models is described. The validation procedure aims to investigate the ability of residential agents to modify their standard consumption patterns by exploiting their flexibility potential in response to the price profile.

This work uses real-world data to construct thermal models and generate stochastic load profiles for a set of residential buildings. The data is related to 11 single-family detached houses, located in the city of Trois-Rivieres, Quebec, Canada. The houses are equipped with electrical baseboards and thermostats for temperature control. The acquisition system records indoor temperature, electrical heating power consumption, and outdoor temperature. The collected data spans four winter months, from January to April 2018. Fig. 4 depicts the conditional density of the power consumption and the difference between the indoor and outdoor temperatures. The measurements have 15-minute sampling intervals. The data allows for constructing linear thermal models of targeted houses. The ridge regression is utilized to determine the coefficients $\boldsymbol{\beta}^i = [\beta_1^i, \beta_2^i, \beta_3^i]$ for the linear model [57],

$$x_{k+1}^i = g(x_k^i, x_k^{\text{out}}, u_{h,k}^i) \tag{12}$$

In addition, the power consumption of energy-extensive appliances other than electric baseboards is considered. This process aims to generate a stochastic aggregate load profile of non-flexible residential appliances [58]. This profile is added to the simulated heating demand. Fig. 5 shows the conditional mean and 95% confidence interval of the weekly load profile for a single house. The data presented is utilized to

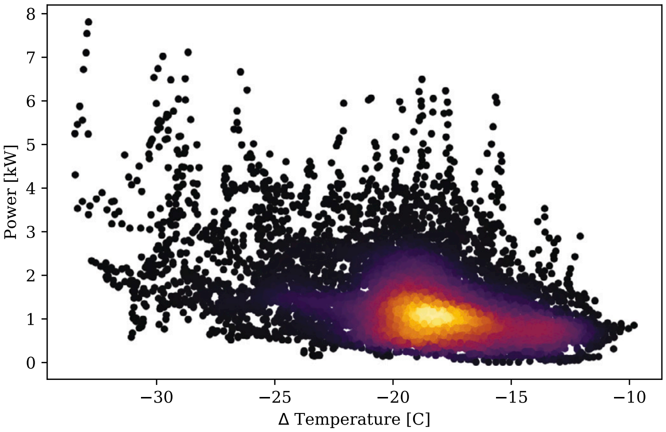


Fig. 4. Distribution of the ESH power consumption and the outdoor temperature for one house in Trois-Rivieres, Quebec.

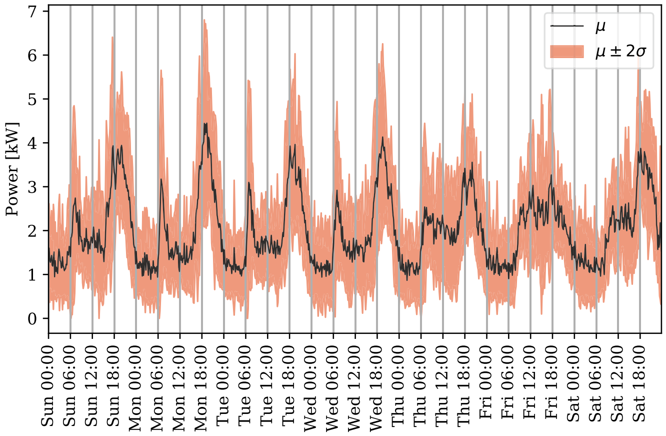


Fig. 5. Average weekly power profile from 8 real houses (space heating load is not included).

obtain the distributions needed to introduce realistic uncertainties for the HEMS optimization simulation process. It should be noted that statistical information from a previous study on temperature preferences in residential buildings is utilized to derive sensible comfort desires for the simulation [59].

For the i^{th} house, the comfortable temperature, x^i_{comf}, is drawn from a discrete distribution as the highest set-point. The generated value is used to compute the household utility function through (10). In this study, the discrete set accounts for four different set-point values obtained by discretizing an empirical distribution over set-point temperatures in Quebec dwellings [59]. The possible values of x^i_{comf} are $[20, 21, 22, 23]$ in degree Celsius [C], and their corresponding probabilities, $P(x^{\text{sp}})$, are $[0.1, 0.3, 0.5, 0.1]$. Besides, the value of the minimum allowed temperature for the same house is generated through $x^i_{\text{min}} = x^i_{\text{comf}} - x^i_{\text{sb}}$, where x^i_{sb} is the set-back value. This quantity is taken randomly from the set $\{1, 2, 3, 4\}$ with $P(x^i_{\text{sb}}) = [0.1, 0.3, 0.4, 0.2]$, calculated by the same manner used for x^i_{comf} [59]. Finally, the value of the parameter δ_{max}, required by the utility function (10), is assumed to be extracted from a log-normal distribution with the expectation, $\mathbb{E}(\delta_{\text{max}})$, and variance, $Var(\delta_{\text{max}})$, equal to 5 and 1, respectively.

4. Results

This section provides the simulation results of the proposed DR mechanism by performing the analysis in three steps. First, validation of the consumption behavior of the residential agents is carried out without the DR mechanisms. Then, the effectiveness of the proposed price generator function for different capacity limits is examined. Finally, the PPO-based RL technique is used to optimize the parameters of the price generator function within the coordination loop to deal with the deviations of the residential agents and maximize the DRA's profits.

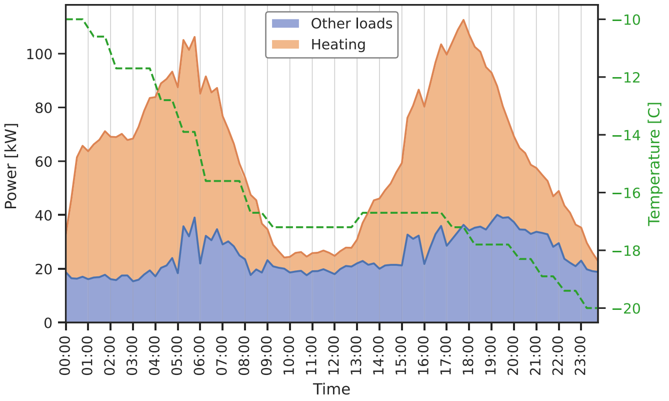


Fig. 6. One-day aggregated power demand without DR.

4.1. The scenario without DR

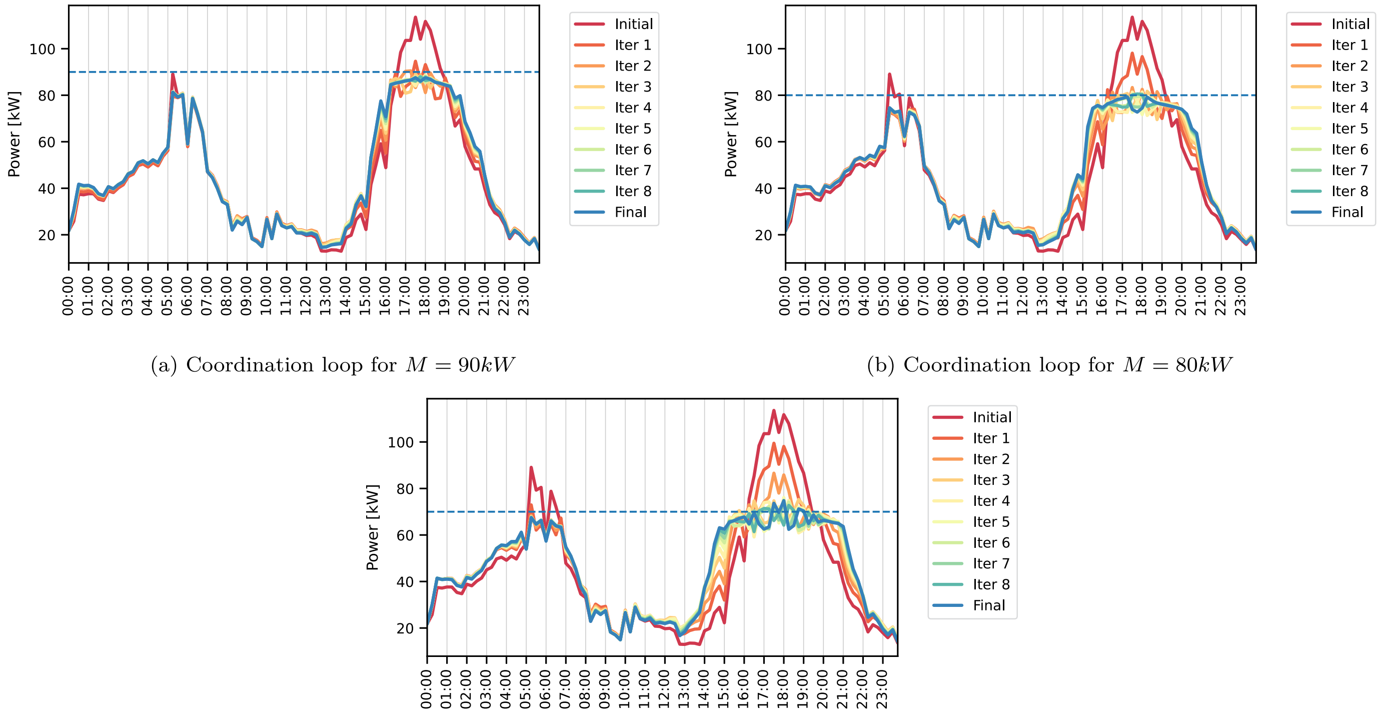
Fig. 6 shows the aggregated consumption profile of a set of 11 simulated buildings during a cold day. The consumption behavior in the figure demonstrates that the models developed are in accordance with the expected power consumption pattern in Quebec's residential sector. Each residential agent performs a model predictive control, meaning they tend to anticipate comfort needs considering the price profile. Therefore, agents will perform actions such as preheating the house before the setpoint temperature changes to x^i_{comf}. From Fig. 6 it can be observed that in the absence of a management mechanism, high peak loads have occurred during morning and evening hours.

4.2. Coordination loop

The performance of the proposed price-based demand response strategy is evaluated utilizing the price generator function proposed in (1). Here, a constrained market is considered, where $\pi_{min} = 0.05\$/kWh$ and $\pi_{max} = 0.20\$/kWh$. The DRA agent starts the coordination loop by establishing a flat price profile. Once aggregating the received response of the users' consumption plan, the DRA agent uses the proposed price generator function (1) to establish the new price policy. This process is performed 10 times before reaching the agreement in the multi-agent system. Fig. 7 shows the results obtained for the capacity constraints $M = 90, 80, 70kW$ for an $\alpha = 5$. The Figure presents the step-by-step interaction between the DRA and the resistive agents. To be more precise, each graph shows the aggregated profiles starting from the users' consumption plan before the DR program's implementation and ending with the consumption profile of the agreement reached. The former is represented in each graph as a red time series and the latter as a blue time series. These results demonstrate that the proposed method allows the translation of a pricing policy into a desired maximum capacity value in a restricted market. Moreover, it can be observed that for higher values of M, residential agents can keep their peak consumption further away from the capacity constraint to exploit further the low price region of the price-generating function. However, as M decreases, this difference is reduced because the users' flexibility starts hitting the limit.

4.3. RL for optimizing DRA pricing strategy

Finally, we evaluate the performance of the proposed PPO-based RL approach in defining the parameters of the price generator function (4)



(a) Coordination loop for $M = 90kW$

(b) Coordination loop for $M = 80kW$

(c) Coordination loop for $M = 70kW$

Fig. 7. Performance analysis of the coordination method for different M values.

during the coordination loop. For this case, the capacity constraint will be established as $M = 75kW$. The RL-based DRA agent seeks to maximize its profit from electricity sales by setting the function's parameters. However, it must also deal with the problem of users' deviations from the consumption plan during its execution. Users try to follow the consumption plan from the agreement as this is the one that maximizes their profit. However, this consumption may deviate from the plan due to possible changes in their activities. Therefore, the DRA agent must be prepared against these changes to avoid being penalized by the DSO. Each RL episode is represented by a coordination loop, which will stop according to criteria based on the change in the percentage of power generation cost reduction with respect to the initial cost and the change in the PAR from one iteration to another. In this case, the coordination will stop when the cost change is less than 0.01%, and the PAR change is less than 0.01. According to the analyses conducted, the proposed criteria are usually met after ten iterations. To better illustrate this, Fig. 8 presents the convergence curve of the coordination loop.

Fig. 9 presents the average curves resulting from the learning process of the DRA agent. The blue curve shows the progression in episodes of the average reward, based on function (5), in red the improvement in PAR at the end of each coordination loop of each episode, and finally in green the aggregator's profit for selling energy using the pricing policy of the agreement. It can be seen that after 600 episodes, the agent improves the reward obtained at the end of the day. In addition, the figure shows how the agent improves its profit per sale of electricity by 35%. At the same time, it offers a reduction of the PAR, demonstrating the performance improvement of the proposed RL method.

Fig. 10 presents a coordination loop between the DRA agent and the residential agents after learning. It can be observed that the implementation of the RL method in the parameter setting of the price generator function enables the DRA agent to utilize the flexibility potential on the residential agent side to improve the aggregate power consumption profile in comparison to the results obtained in Fig. 7. A remarkable point is the amount of electricity consumption shifted from the peaks to the valley. This type of behavior is due to the nature of the controllable load of the residential agents. In houses with electric space heating systems exposed to winter temperatures, the set-point profiles have a significant incidence on initial consumption peaks. For the control mechanisms, these values are used to determine the thermal preference profiles of residential users. This means that for higher set-point periods, the residential agent assumes that a greater need for thermal comfort is requested. Therefore, during lower values, these periods are used to give the residential agent the freedom to control the indoor temperature freely. This means that internally, the house must be preheated to a higher temperature than the higher set-point so that the need for heating is reduced during peak consumption. Because of this preheating, a greater increase in consumption during the valley is likely to be found to meet thermal comfort needs during the peaks.

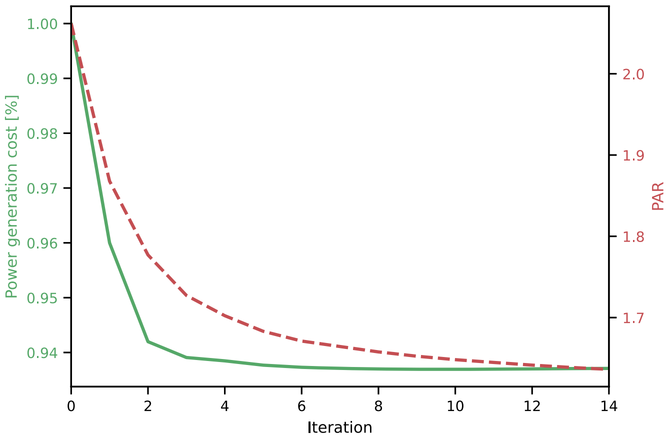


Fig. 8. Power generation cost percentage and PAR curves during coordination loop.

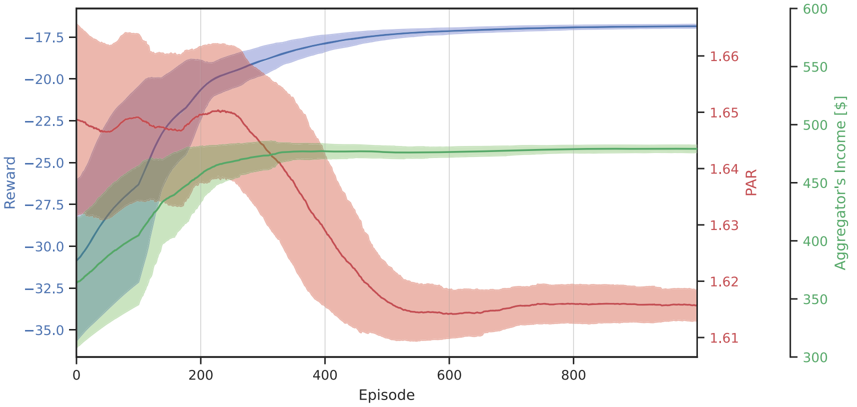


Fig. 9. Analysis of DRA agent performance during the learning process.

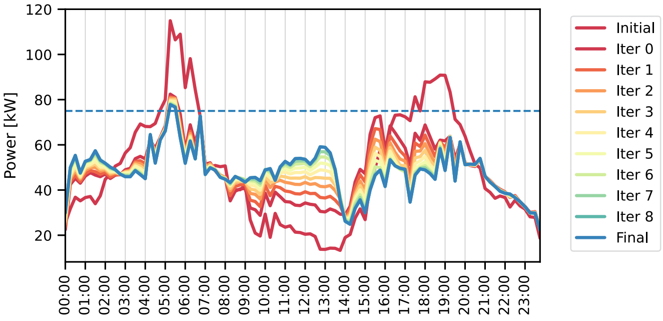


Fig. 10. Coordination loop after RL learning process with $M = 75kW$.

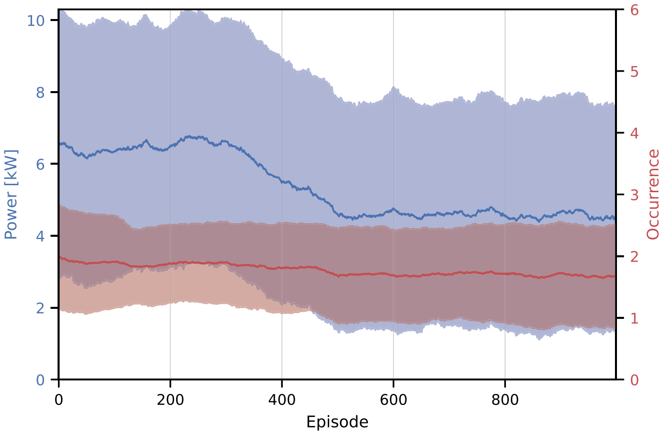


Fig. 11. Analysis of average capacity constraint overruns.

In terms of deviation, Fig. 11 The Figure presents the results related to the difference between the established capacity limit and the maximum peak consumption of the users after the execution of their consumption plan. For this purpose, the final calculation of the reward function is performed after the execution of the consumption plans, i.e., the calculation of the reward is made using the consumption profile $\hat{u}_k$. Considering those deviations in the plan, the blue curve represents the average spread of the differences between the maximum peak consumption during the 24 hours and the capacity limit. In addition, the red curve indicates the occurrence of exceeding this limit, measured in a number of timesteps encountered in excess of the M limit. These results illustrate that the DRA agent maintains a trend in decreasing the average occurrence of exceeding the capacity constraint. In addition, the figure also shows that the agent decreases the power difference between the constraint and the consumption peak throughout the learning process. Finally, Fig. 12 presents the profile of the aggregate consumption plan and pricing policy stipulated in the agreement, and the actual aggregate consumption of the houses after the 600 episodes. The proposed method demonstrates the effectiveness of the proposed strategy in dealing with uncertainty arising from deviations from the consumption plan of residential agents. As it is represented, the DRA even tries to accept a slight deviation from the consumption plan of the agreement in order to use these deviations to its advantage in the execution. This in order to obtain a higher profit from the sale of energy. However, this type of behavior could be avoided by adjusting the values of w_1 and w_2 in equation (3).

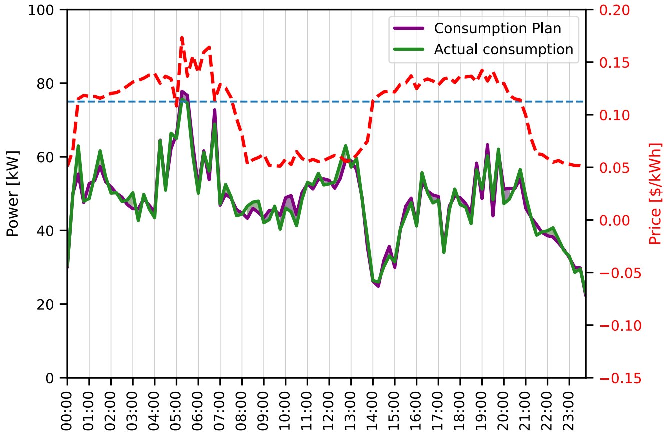
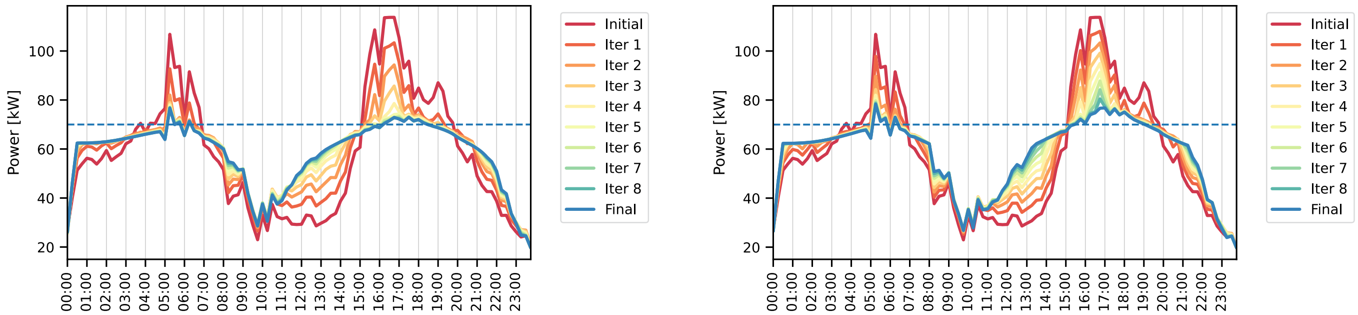


Fig. 12. Difference between actual aggregate consumption and the consumption plan of the agreement under the established price profile.

4.4. Performance comparison

To determine the effectiveness of the selected approach, a performance comparison was made for both the proposed price generator function and the implemented RL mechanism. First, we compare the price function (4) with a standard piece-wise linear function. This new function was constructed based on the derivative of our sigmoid function to ensure an approximate shape between them. Another winter day was selected randomly to verify the performance of the proposed generator in exploiting the flexibility potential of a set of residential customers. Fig. 13 provides a performance comparison within the coordination loop, for $M = 70kW$. The results illustrate that the proposed sigmoid function (1) is able to exploit, in a superior manner, the flexibility potentials of the residential agents, considering the same environmental conditions. This can be noticed by comparing overruns of the



(a) Coordination loop the proposed sigmoid function.

(b) Coordination loop for a piece-wise linear function.

Fig. 13. Performance analysis of different price generator functions.

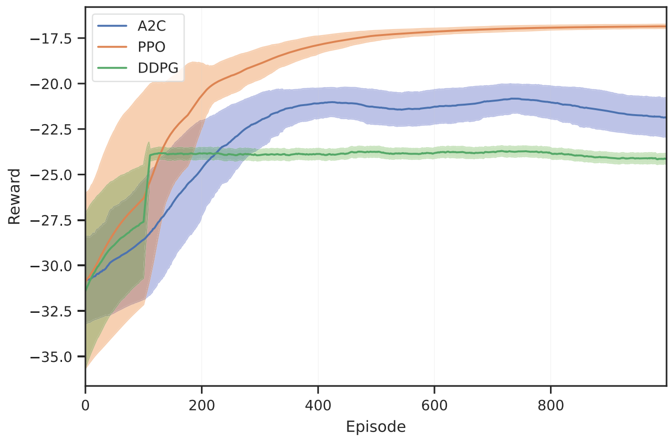


Fig. 14. Comparative performance of our PPO mechanism with A2C and DDPG.

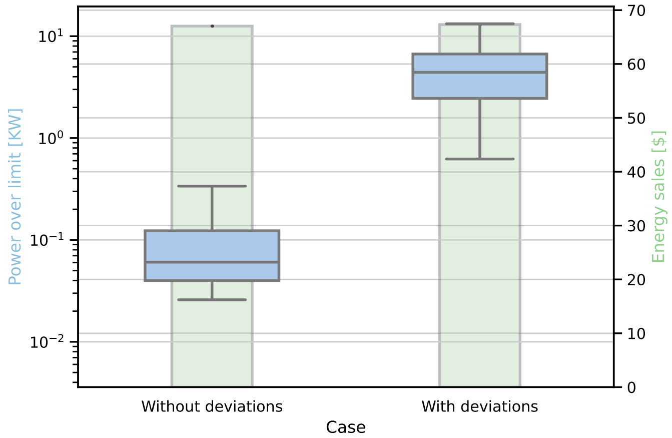


Fig. 15. Performance analysis related to the consideration of users' deviations from consumption plans.

capacity limitation M. For instance, in terms of the number of overruns, the sigmoid function outperforms the piece-wise linear function by achieving 41% fewer overruns at the end of the coordination loop. Furthermore, the power consumption over M is higher in the piece-wise linear function by 75%, evidencing the significant differences in terms of flexibility exploitation.

Taking into account the performance of the RL algorithm, the selected PPO mechanism was compared with the popular Advantage-Actor-critic (A2C) and Deep Deterministic Policy Gradient (DDPG) methods. Fig. 14 provides the curves of the progression in iterations of the average reward, based on function (5). The results demonstrate that the selected approach provides better efficiency in dealing with the uncertainty of the scenario encountered. According to this Figure, the PPO and A2C algorithms are able to obtain better results than DDPG. Furthermore, the PPO mechanism converges to a solution that provides a reward 38% higher than the A2C method, meaning that by implementing the PPO algorithm, the DRA agent will be able to capitalize its effort in terms of higher profits from energy selling and DSO reward received.

Finally, to better illustrate the performance of the proposed method, a last comparison is performed, taking into account the uncertainty in the behavior of residential users. Fig. 15 provides a comparative result after the training process during 20 days of the winter season. It is possible to verify that the average results are almost the same in terms of DRA's profit from energy sales. However, considering overruns of the capacity limit, there exists a significant difference as in the case without the uncertainty, the average cumulative daily power over the limit is $0.05kW$, but in the case where the deviations are considered, the accumulated power is around $4kW$. This can translate to a better exploitation of the DSO's reward and a higher DRA's profit when this uncertainty is not considered.

5. Discussions and future prospects

The optimal generation of pricing policies has been a critical aspect in implementing price-based DR programs. Moreover, the consideration of existing regulations would be an important issue in the implementation of these programs. These regulations define limits on price sales per energy unit, creating new constraints for the optimization problems existing in the literature and affecting the optimality of their solutions. Another key aspect is the goal of these DR mechanisms in the residential sector. Their goal is to exploit their flexibility potential to reduce consumption peaks. However, implementing such strategies can result in imbalances and losses in the power grid if the system's real needs are not considered [57]. In this regard, some studies have been conducted in the literature considering pricing policies where capacity limits are established [32], especially in the presence of electric vehicles [60]. However, integrating these capacity limitations, taking into account other sources of flexibility from the residential sector, needs to be further explored. This is an important point as Smart Energy Systems are focused on merging the electricity, heating, and transport sectors with storage options to foster the adaptability required for accommodating significant amounts of fluctuating renewable energy [61]. This clearly expresses the need for integrating electric heating systems with new flexibility sources like battery electric vehicles in the same capacity-constrained scenario. Therefore, the aforementioned highlights the importance of developing new strategies, such as the one presented in this paper, to facilitate the future integration of heating systems with emerging technologies in residential smart energy systems.

The traditional fixed-rate pricing schemes have been widely implemented around the world. However, the increase in price volatility has made the retailers migrate to more dynamic pricing strategies like Time-of-Use programs. This means that we are at a stage where hourly rates are becoming a standard, and therefore, it is expected that the rate time resolution will soon drop by 15 minutes, as is the case in Europe. [32]. For this reason, it is necessary to develop dynamic pricing mechanisms, such as the one presented in this paper, to allow the management and optimization of residential consumption in these evolving scenarios. In particular, the consideration of the energy consumption of the heating sector in this type of scheme facilitates the intended energy transition and contributes to limiting the need for new infrastructures, as shown in [62].

In this sense, it is important to define strategies that allow users to coordinate through these pricing policies. This represents a great benefit for entities such as the DSO, as presented in [29]. In this paper, the authors propose a dynamic pricing mechanism that significantly reduces consumption peaks. This is achieved through a coordination loop in which pricing policies proportional to the aggregate consumption profile are used, allowing users' privacy to be respected. However, price limits are not considered for generating the policies, hindering the possibility of their implementation under the existing regulations in the energy markets. This can also lead to significant decreases in energy sales profits, as shown in [37]. For this reason, the approach proposed in this work considered the utilization of a dynamic price generator function by a DRA to improve the ideas presented in [29]. This function performs a monotonic transformation of the aggregate consumption profile, taking into account price constraints and capacity limits, allowing the achievement of a reduction in peak consumption in a more controllable manner. As a result, the way in which user flexibility is managed enables the opportunity to offer capacity services to the DSO, and highlights the benefits of exploiting the flexibility potentials of heating systems for the system.

The performance of this function is compared with a piece-wise linear function, demonstrating how the proposed sigmoid-based function provides better management of the residential flexibility by accomplishing significant results in terms of capacity overruns. However, it is not an easy task to determine the correct parameter settings of this function, as any information from the demand side is known by the DRA. Moreover, users can deviate from their stipulated consumption plans during run time due to external variables or unexpected events that may affect non-controllable load consumption. For this reason, a Deep-RL mechanism is proposed to handle the uncertainties related to the lack of this information. The results evidence that the RL-based DRA is able to set the parameters of the proposed price generator function properly in order to guarantee the capacity limit and price constraints while maximizing its profit for selling energy. This significant achievement can contribute to the smart energy system transition by reducing the electricity demand consciously, which indirectly influences power generation. To illustrate, this could mean a reduction of biomass consumption, increasing the feasibility of carrying out energy transition strategies such as the one presented in [63].

In order to improve the obtained results, further considerations must be taken into account. For instance, the integration of energy storage systems may be very beneficial, as these systems can help with the absorption of energy consumption deviations from the demand side. This can allow a better performance of the mechanism proposed in terms of players' profits and increase flexibility opportunities within smart energy systems. Furthermore, the integration of electric vehicles must be prospectively evaluated to analyze the effect of capacity limitations for electric vehicle charging on the management of the heating sector. The implementation of the proposed DR program, based on dynamic pricing, should be carried out to evaluate the effect on demand response under the management of these two different types of loads.

6. Conclusions

In this paper, a price-based DR program is proposed that incorporates power capacity and market constraints to coordinate a set of residential agents. For this purpose, a price generator function is proposed, considering existing market regulations that limit energy sales prices. This function allows translating the maximum desirable capacity into a pricing policy through a coordination loop in a Stackelberg game-theoretic framework, obtaining a mechanism that allows exploiting residential flexibility in a more controlled way. The price generator function performance is demonstrated through a comparison against a linear piece-wise function, evidencing 41% fewer overrun and a power consumption over the capacity limit 75% lower at the end of the coordination loop. Furthermore, an RL-based DRA agent utilizes this price generator to define pricing policies that maximize its profit in the constrained proposed scenario, where the DRA needs to deal with deviations from users' stipulated consumption plans. The proposed strategy was able to exploit residential agents' flexibility, adjusting the parameters of the price generator function within the coordination loop. Moreover, the proposed approach evidences the viability of exploiting the flexibility potentials of electric space heating systems from the residential sector, in such scenarios where capacity limitations are required from the DSO. The simulation results demonstrated that the proposed DR strategy improved DRA's profits by 35% while dealing with residential agents' deviations. The comparative study displayed the superiority of the proposed price-based DR program and the adopted PPO-based RL technique converging to a solution that provides a reward 38% higher for the DRA than the well-known A2C and DDPG methods.

CRediT authorship contribution statement

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

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.

Data availability

The data that has been used is confidential.

Acknowledgement

This work was supported in part by the Laboratoire des technologies de l'energie (LTE) d'Hydro-Quebec, the Natural Science and Engineering Research Council of Canada and the Foundation of Université du Québec à Trois-Rivières.

Appendix A. PPO algorithm

Procedure for the implementation of the proposed dynamic pricing mechanism based on PPO.

Algorithm 1: PPO algorithm.

```
DSO communicates the desirable capacity limit $M$.
The DRA asks residential agents for their stipulated consumption plan
  under a constant price.
DRA determines the initial state $s_0$.
for $t = 0, 1, 2, \ldots$ do
    Define the action $a_t = \{\eta, \alpha\}$. (Transformation of Price function (4)
      defined by the aggregator agent)
    Each Residential agent solves its own optimization problem
      expressed in (11).
    Get the normalized state $s_t$. (Aggregated residential agents' response)
    Calculate rewards-to-go $R_t$ based on (5).
    Collect the set of partial trajectories $\{(s_t, a_t, R_t, s_t + 1)\}$ on policy
      $\phi_t = \phi_{\theta_t}(a_t, s_t)$.
    Estimate advantage $\hat{A}_t$.
    if $t$ mod $T = 0$ then
        Compute policy update by means of (6):
        $\theta_{t+1} = \arg\max_{\theta} \sum_{j=0}^{T} J(\theta)$
        via stochastic gradient ascent with Adam [48].
    end
end
```

References

[1] Kumar RS, Raghav LP, Raju DK, Singh AR. Impact of multiple demand side management programs on the optimal operation of grid-connected microgrids. Appl Energy 2021;301:117466.

[2] Fotouhi Ghazvini MA, Soares J, Abrishambaf O, Castro R, Vale Z. Demand response implementation in smart households. Energy Build 2017;143:129–48. https://doi.org/10.1016/j.enbuild.2017.03.020.

[3] Asadinejad A, Tomsovic K. Optimal use of incentive and price based demand response to reduce costs and price volatility. Electr Power Syst Res 2017;144:215–23. https://doi.org/10.1016/j.epsr.2016.12.012.

[4] Vuelvas J, Ruiz F. A novel incentive-based demand response model for Cournot competition in electricity markets. Energy Syst 2019;10(1):95–112. https://doi.org/10.1007/s12667-018-0271-2.

[5] Palensky P, Dietrich D. Demand side management: demand response, intelligent energy systems, and smart loads. IEEE Trans Ind Inform 2011;7(3):381–8. https://doi.org/10.1109/TII.2011.2158841.

[6] Hu Q, Li F, Fang X, Bai L. A framework of residential demand aggregation with financial incentives. IEEE Trans Smart Grid 2018;9(1):497–505. https://doi.org/10.1109/TSG.2016.2631083.

[7] D'hulst R, Labeeuw W, Beusen B, Claessens S, Deconinck G, Vanthournout K. Demand response flexibility and flexibility potential of residential smart appliances: experiences from large pilot test in Belgium. Appl Energy 2015;155:79–90. https://doi.org/10.1016/j.apenergy.2015.05.101.

[8] Burger S, Chaves-Ávila JP, Batlle C, Pérez-Arriaga IJ. The value of aggregators in electricity systems. https://energy.mit.edu/wp-content/uploads/2016/01/CEEPR_WP_2016-001.pdf, 2016.

[9] Yuan Z-P, Li P, Li Z-L, Xia J. Data-driven risk-adjusted robust energy management for microgrids integrating demand response aggregator and renewable energies. IEEE Trans Smart Grid 2023;14(1):365–77. https://doi.org/10.1109/TSG.2022.3193226.

[10] Olivella-Rosell P, Lloret-Gallego P, Munné-Collado Í, Villafafila-Robles R, Sumper A, Ottessen SØ, et al. Local flexibility market design for aggregators providing multiple flexibility services at distribution network level. Energies 2018;11(4):1–19. https://doi.org/10.3390/en11040822.

[11] Gjorgievski VZ, Markovska N, Abazi A, Duić N. The potential of power-to-heat demand response to improve the flexibility of the energy system: an empirical review. Renew Sustain Energy Rev 2021;138:110489. https://doi.org/10.1016/j.rser.2020.110489.

[12] Hosseini S, Kelouwani S, Agbossou K, Cardenas A, Henao N. A semi-synthetic dataset development tool for household energy consumption analysis. In: 2017 IEEE international conference on industrial technology (ICIT); 2017. p. 564–9.

[13] Duman AC, Erden HS, Gönül Ömer, Güler Önder. A home energy management system with an integrated smart thermostat for demand response in smart grids. Sustain Cities Soc 2021;65:102639. https://doi.org/10.1016/j.scs.2020.102639.

[14] Yan X, Ozturk Y, Hu Z, Song Y. A review on price-driven residential demand response. Renew Sustain Energy Rev 2018;96:411–9.

[15] Yassine A. Implementation challenges of automatic demand response for households in smart grids. In: 2016 3rd international conference on renewable energies for developing countries (REDEC); 2016. p. 1–6.

[16] Silva C, Faria P, Vale Z, Corchado J. Demand response performance and uncertainty: a systematic literature review. Energy Strategy Rev 2022;41:100857. https://doi.org/10.1016/j.esr.2022.100857.

[17] Dominguez J, Parrado-Duque A, Montoya OD, Henao N, Campillo J, Agbossou K. Techno-economic feasibility of a trust and grid-aware coordination scheme. In: 2023 IEEE Texas power and energy conference (TPEC); 2023. p. 1–5.

[18] O'Connell N, Pinson P, Madsen H, O'Malley M. Benefits and challenges of electrical demand response: a critical review. Renew Sustain Energy Rev 2014;39:686–99. https://doi.org/10.1016/j.rser.2014.07.098.

[19] Venizelou V, Philippou N, Hadjipanayi M, Makrides G, Efthymiou V, Georghiou GE. Development of a novel time-of-use tariff algorithm for residential prosumer price-based demand side management. Energy 2018;142:633–46.

[20] Rasheed MB, Qureshi MA, Javaid N, Alquthami T. Dynamic pricing mechanism with the integration of renewable energy source in smart grid. IEEE Access 2020;8:16876–92. https://doi.org/10.1109/ACCESS.2020.2967798.

[21] Ohannessian MI, Roozbehani M, Materassi D, Dahleh MA. Dynamic estimation of the price-response of deadline-constrained electric loads under threshold policies. In: 2014 American control conference; 2014. p. 2798–803.

[22] Zhang D, Zhu H, Zhang H, Goh HH, Liu H, Wu T. Multi-objective optimization for smart integrated energy system considering demand responses and dynamic prices. IEEE Trans Smart Grid 2022;13(2):1100–12. https://doi.org/10.1109/TSG.2021.3128547.

[23] Jia L, Tong L. Dynamic pricing and distributed energy management for demand response. IEEE Trans Smart Grid 2016;7(2):1128–36. https://doi.org/10.1109/TSG.2016.2515641.

[24] Taherian H, Aghaebrahimi MR, Baringo L, Goldani SR. Optimal dynamic pricing for an electricity retailer in the price-responsive environment of smart grid. Int J Electr Power Energy Syst 2021;130:107004. https://doi.org/10.1016/j.ijepes.2021.107004.

[25] Guo Z, Xu W, Yan Y, Sun M. How to realize the power demand side actively matching the supply side?——a virtual real-time electricity prices optimization model based on credit mechanism. Appl Energy 2023;343:121223.

[26] Hong Q, Meng F, Liu J, Bo R. A bilevel game-theoretic decision-making framework for strategic retailers in both local and wholesale electricity markets. Appl Energy 2023;330:120311.

[27] Aguiar N, Dubey A, Gupta V. Network-constrained Stackelberg game for pricing demand flexibility in power distribution systems. IEEE Trans Smart Grid 2021;12(5):4049–58. https://doi.org/10.1109/TSG.2021.3078905.

[28] Collins LD, Middleton RH. Distributed demand peak reduction with non-cooperative players and minimal communication. IEEE Trans Smart Grid 2019;10(1):153–62. https://doi.org/10.1109/TSG.2017.2734113.

[29] Nguyen HK, Song JB, Han Z. Distributed demand side management with energy storage in smart grid. IEEE Trans Parallel Distrib Syst 2015;26(12):3346–57. https://doi.org/10.1109/TPDS.2014.2372781.

[30] Margellos K, Oren S. Capacity controlled demand side management: a stochastic pricing analysis. IEEE Trans Power Syst 2016;31(1):706–17. https://doi.org/10.1109/TPWRS.2015.2406813.

[31] Heinrich C, Ziras C, Jensen TV, Bindner HW, Kazempour J. A local flexibility market mechanism with capacity limitation services. Energy Policy 2021;156:112335. https://doi.org/10.1016/j.enpol.2021.112335.

[32] Crowley B, Kazempour J, Mitridati L. Dynamic pricing in an energy community providing capacity limitation services. arXiv:2309.05363, 2023.

[33] Yun L, Wang D, Li L. Explainable multi-agent deep reinforcement learning for real-time demand response towards sustainable manufacturing. Appl Energy 2023;347:121324.

[34] Lu R, Hong SH, Zhang X. A dynamic pricing demand response algorithm for smart grid: reinforcement learning approach. Appl Energy 2018;220:220–30. https://doi.org/10.1016/j.apenergy.2018.03.072.

[35] Kim B-G, Zhang Y, van der Schaar M, Lee J-W. Dynamic pricing for smart grid with reinforcement learning. In: 2014 IEEE conference on computer communications workshops (INFOCOM WKSHPS); 2014. p. 640–5.

[36] Du Y, Li F. Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning. IEEE Trans Smart Grid 2020;11(2):1066–76. https://doi.org/10.1109/TSG.2019.2930299.

[37] Fraija A, Agbossou K, Henao N, Kelouwani S, Fournier M, Hosseini SS. A discount-based time-of-use electricity pricing strategy for demand response with minimum information using reinforcement learning. IEEE Access 2022;10:54018–28. https://doi.org/10.1109/ACCESS.2022.3175839.

[38] Liu Y, Zuo K, Liu XA, Liu J, Kennedy JM. Dynamic pricing for decentralized energy trading in micro-grids. Appl Energy 2018;228:689–99. https://doi.org/10.1016/j.apenergy.2018.06.124.

[39] Huang H, Ning Y, Jiang Y, Tang Z, Qian Y, Zhang X. Dynamic pricing optimization for commercial subcontracting power suppliers engaging demand response considering building virtual energy storage. Front Energy Res 2024;11. https://doi.org/10.3389/fenrg.2023.1329227.

[40] Wu C, Mohsenian-Rad H, Huang J, Wang AY. Demand side management for wind power integration in microgrid using dynamic potential game theory. In: 2011 IEEE GLOBECOM workshops (GC Wkshps); 2011. p. 1199–204.

[41] Lee S, Choi D-H. Dynamic pricing and energy management for profit maximization in multiple smart electric vehicle charging stations: a privacy-preserving deep reinforcement learning approach. Appl Energy 2021;304:117754.

[42] Huang M, Liu C, He X, Ma L, Lu Z, Su H. Reinforcement learning-based control for nonlinear discrete-time systems with unknown control directions and control constraints. Neurocomputing 2020;402:50–65.

[43] Vázquez-Canteli JR, Nagy Z. Reinforcement learning for demand response: a review of algorithms and modeling techniques. Appl Energy 2019;235:1072–89. https://doi.org/10.1016/j.apenergy.2018.11.002.

[44] Coraci D, Brandi S, Hong T, Capozzoli A. Online transfer learning strategy for enhancing the scalability and deployment of deep reinforcement learning control in smart buildings. Appl Energy 2023;333:120598. https://doi.org/10.1016/j.apenergy.2022.120598.

[45] Wiering M, van Otterlo M. Reinforcement learning: state-of-the-art, adaptation, learning, and optimization. Springer Berlin Heidelberg; 2012. https://books.google.ca/books?id=T4wovQEACAAJ.

[46] Wan Y, Qin J, Yu X, Yang T, Kang Y. Price-based residential demand response management in smart grids: a reinforcement learning-based approach. IEEE/CAA J Autom Sin 2022;9(1):123–34. https://doi.org/10.1109/JAS.2021.1004287.

[47] Wang Y, Qiu D, Sun M, Strbac G, Gao Z. Secure energy management of multi-energy microgrid: a physical-informed safe reinforcement learning approach. Appl Energy 2023;335:120759.

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

[49] Deng R, Yang Z, Chen J, Chow MY. Load scheduling with price uncertainty and temporally-coupled constraints in smart grids. IEEE Trans Power Syst 2014;29(6):2823–34. https://doi.org/10.1109/TPWRS.2014.2311127.

[50] Olesen BW, Brager GS. A better way to predict comfort: the new ashrae standard 55-2004; 2004.

[51] Jose JAO. A review of general and local thermal comfort models for controlling indoor ambiences. In: Kumar A, editor. Air quality. Rijeka: IntechOpen; 2010. Ch. 14.

[52] Stavrakas V, Flamos A. A modular high-resolution demand-side management model to quantify benefits of demand-flexibility in the residential sector. Energy Convers Manag 2020;205:112339. https://doi.org/10.1016/j.enconman.2019.112339.

[53] Nematirad R, Ardehali MM, Khorsandi A. Optimization of residential demand response program cost with consideration for occupants thermal comfort and privacy. arXiv:2305.08077, 2023.

[54] Dong Z, Zhang X, Li Y, Strbac G. Values of coordinated residential space heating in demand response provision. Appl Energy 2023;330:120353.

[55] Scutari G, Palomar DP, Facchinei F, Pang J-S. Monotone games for cognitive radio systems. London: Springer London; 2012. p. 83–112.

[56] Chen Y, Lin WS, Han F, Yang Y-H, Safar Z, Liu KJR. Incentive compatible demand response games for distributed load prediction in smart grids. APSIPA Trans Signal Inf Process 2014;3:e9. https://doi.org/10.1017/ATSIP.2014.8.

[57] Domínguez-Jiménez J, Henao N, Agbossou K, Parrado A, Campillo J, Nagarsheth SH. A stochastic approach to integrating electrical thermal storage in distributed demand response for nordic communities with wind power generation. IEEE Open J Ind Appl 2023;4:121–38. https://doi.org/10.1109/OJIA.2023.3264651.

[58] Toquica D, Agbossou K, Henao N, Malhamé R, Kelouwani S, Amara F. Prevision and planning for residential agents in a transactive energy environment. Smart Energy 2021;2:100019. https://doi.org/10.1016/j.segy.2021.100019.

[59] Henao N, Fournier M, Kelouwani S. 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. p. 17–25.

[60] Roy P, Ilka R, He J, Liao Y, Cramer AM, Mccann J, et al. Impact of electric vehicle charging on power distribution systems: a case study of the grid in western Kentucky. IEEE Access 2023;11:49002–23. https://doi.org/10.1109/ACCESS.2023.3276928.

[61] Mathiesen B, Lund H, Connolly D, Wenzel H, Østergaard P, Möller B, et al. Smart energy systems for coherent 100% renewable energy and transport solutions. Appl Energy 2015;145:139–54. https://doi.org/10.1016/j.apenergy.2015.01.075.

[62] Connolly D, Lund H, Mathiesen B. Smart energy Europe: the technical and economic impact of one potential 100% renewable energy scenario for the European Union. Renew Sustain Energy Rev 2016;60:1634–53. https://doi.org/10.1016/j.rser.2016.02.025.

[63] Hansen K, Mathiesen BV, Skov IR. Full energy system transition towards 100% renewable energy in Germany in 2050. Renew Sustain Energy Rev 2019;102:1–13. https://doi.org/10.1016/j.rser.2018.11.038.

3.3 Multi-aggregator system

3.3.1 Background

With the multi-agent environment set for the interactions between a DRA and a set of houses and the definition of a price generator function for parameterizing the dynamic price mechanism, it is now time to evaluate this policy generation from the system point of view. As was mentioned before, individual solutions from different DRAs do not guarantee that an optimal good solution will be achieved. This work delves deeper into dynamic pricing with multiple DRAs based on multi-agent reinforcement learning (MARL), as each aggregator will continue interacting with their own set of houses.

As one of our goals, this approach ensures customers' privacy while they generate optimal responses that minimize their costs and maximize their benefits. However, to make the DRAs coordinate, it is necessary to implement a fair reward allocation mechanism from the upper level, according to their contributions to the system's objective. This is where Shapley value (SV), a concept from cooperative game theory, comes into play. Each DRA will receive a reward from the DSO based on its contribution to the global objective through SV calculation. The integration of SV provides a fair and mathematically grounded framework for distributing the benefit of cooperating among the DRA agents. Furthermore, this work also demonstrates that the assessment of rewards based on their marginal impact on the overall system expedites the performance of the MARL architecture, improving the agents' understanding of the impact of their actions on the environment [67].

3.3.2 Methodology

To evaluate the mechanism, building the MARL environment considering the multi-aggregator system is necessary. Based on the previous development of this thesis, a set of DRAs is established, each one interacting with a set of houses with different cardinality. Furthermore, each customer will have their comfort preferences, affecting their response to the transactive signals generated by the DRA. As a result, each aggregator must learn their strategies in a decentralize-training-decentralize-execution (DTDE) MARL architecture [68].

Each DRA agent will use the price generator function to offer discounts to incentivize customers to modify their consumption patterns. At the beginning of the day, a coordination loop is performed between each DRA and its set of customers until an agreement is reached. The results of this interaction will establish the dynamic price tariff and the consumption plan for the customers during the following 24 hours. At the end of the day, the DSO will determine the marginal contribution in accomplishing the global objective of the system to each DRA, and based on this, the aggregators' rewards are defined. Figure 3.2, provides a representation of the interaction between the different system players.

The simulation is performed considering the environmental condition of a winter day in Quebec. The decision-making process for each aggregator will be defined by implementing the Independent PPO (IPPO) MARL mechanism. Then, a performance analysis is conducted by applying the IPPO method with and without the SV-based reward-sharing mechanism.

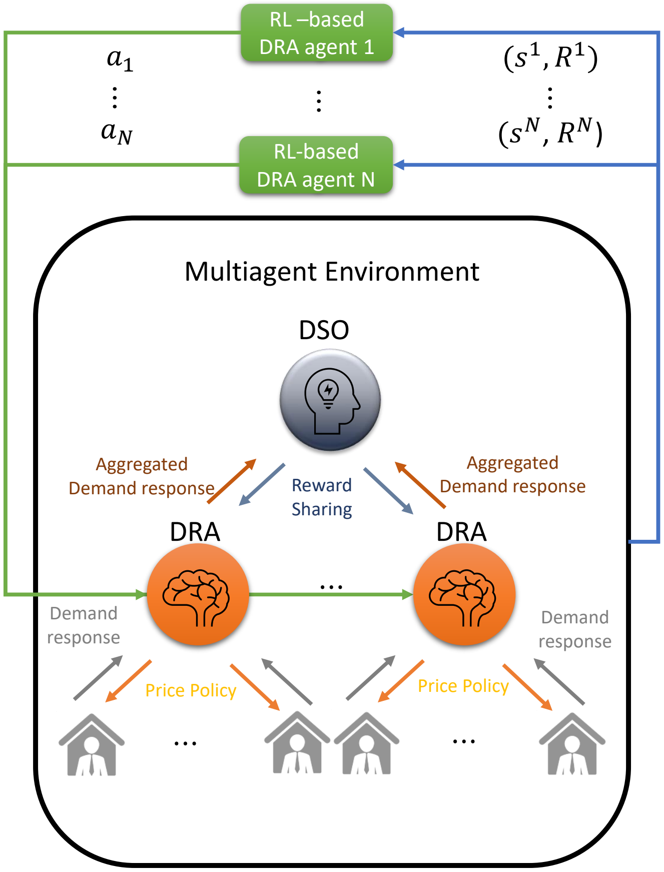


Figure 3.2 Multi-agent system interaction.

3.3.3 Outcomes

This work proposes a cooperative price-based demand response mechanism for a multi-aggregator system based on MARL and an SV-based reward-sharing mechanism. As the DRAs establish dynamic pricing discounts in an iterative process, the customers can adapt their consumption profiles to gain advantages of these discounts. This strategy creates a win-win approach, as the residential users can exploit the flexibility of their controllable loads to reduce their bills, while the DRAs can offer this flexibility to the DSO to reduce the system's aggregated peak demand. By means of the MARL strategy, the DRAs make a trade-off between reducing their profit by offering discounts to the houses for exploiting their flexibility potential and the reward that the DSO offers them for the effort made. The results presented demonstrate a significant PAR reduction in the total power demand. Furthermore, the importance of implementing the SV-based reward-sharing mechanism is shown in terms of improving the solution optimization and reducing the IPPO convergence time. This is because the use of the calculation of the marginal contribution of each aggregator agent helps to deal with the major problem of MARL techniques, which is the non-stationarity of the environment.

Highlights

Cooperative Price-based Demand Response Program for Multiple Aggregators based on Multi-agent Reinforcement Learning and Shapley-Value

Alejandro Fraija, Nilson Henao, Kodjo Agbossou, Sousso Kelouwani, Michaël Fournier

- A Cooperative multi-aggregator system is proposed for a set of DRA agents.
- A MARL architecture is proposed to determine dynamic pricing strategies.
- A fair reward-sharing mechanism is used to estimate the gain of RL-based DRA agents.
- Results evidence the coordination of DRAs to achieve a global system goal.

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*

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%.

Keywords: Demand response, demand response aggregator, dynamic pricing, multi-agent reinforcement learning, Shapley-value.

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

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

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

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 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, 24, 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]. Considering these limitations, RL approaches have emerged as a valuable option to deal with problems related to the optimal price policy generation process. For instance, in [27], a Q-learning method was adopted to decide the retail electricity price, considering service provider and customers profit, without requiring the full knowledge of the system dynamics and uncertainties. In [28], 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.

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 [29]. This is where DRA effectively facilitates customer participation by working in a more customer-oriented manner [30]. 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 [31], 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 [32], 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 [33], active voltage control is proposed, based on Dec-POMDP, to enable real-world applications of MARL algorithms in power systems. Authors in [34] implemented a MARL approach to controlling a complex system of production resources, battery storage, electricity self-supply, and short-term market trading. In [35], authors demonstrate the value of MARL mechanism, which can quickly optimize thermostatically controlled loads performance by applying collaborative multi-agent decision-making processes. In [36], 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 [37] 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. However, these previous approaches have not considered fairness in the reward allocation process for each RL-based agent. Proposing MARL as a pricing approach for multi-aggregator systems makes determining a fair incentive allocation strategy necessary. In [38], authors demonstrated that combining the DR programs with SV helps retailers assure profitability and also enhances user participation. Authors in [39, 40] 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.

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

Ref	**DR Mechanism**	**Solution Method**	**Multiple aggregators**	**Price policy optimization**	**Reward sharing mechanism**
[18]	Dynamic pricing	Price responsive modeling	✗	✓	✗
[19]	Dynamic pricing	Bi-level, meta-heuristic	✗	✓	✗
[20]	Real-time pricing	Stochastic optimization	✗	✓	✗
[21]	Dynamic pricing	Multi-objective optimization	✗	✓	✗
[22]	Time-of-Use	Multi-objective optimization	✗	✓	✗
[23]	Dynamic pricing	Game-theoretic model	✗	✓	✗
[24, 25]	Time-of-Use	Game-theoretic model	✗	✓	✗
[27, 28]	Dynamic pricing	RL	✗	✓	✗
[9]	Load following signals	H-ADMM	✓	✗	✗
[28]	Diverse compensation price	Game-theoretic model	✓	✓	✗
[36]	Incentive-based	MARL	✗	✗	✗
[37]	Dynamic pricing	MARL	✗	✓	✗
Proposed work	Dynamic pricing	MARL	✓	✓	✓

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.

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

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 Figure 1, the DSO rewards each DRA for contributing to the peak shaving objective in a day-ahead scheme. In response, the DRA stipulates price

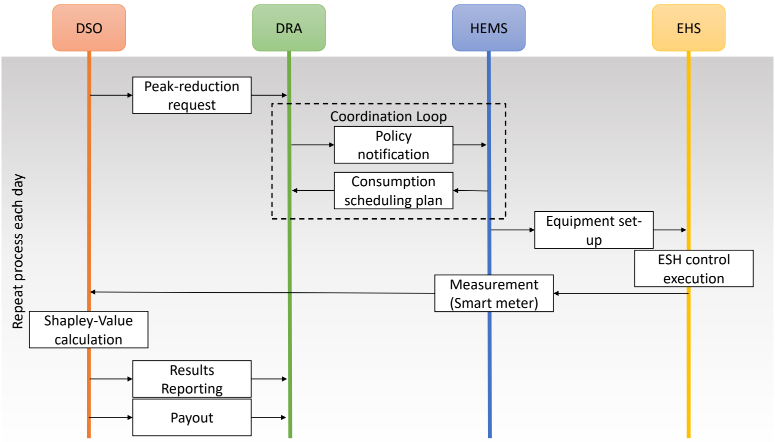


Figure 1: Automatic DR sequence for the multi-aggregator system.

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. Figure 2, provides a representation of the interaction between the different actors of the proposed scenario. As seen in Figures 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 [41].

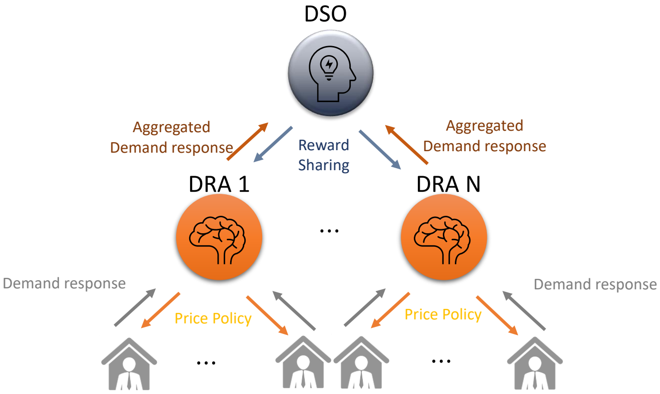


Figure 2: Interaction between market participants in the DR program.

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 $\boldsymbol{\lambda}$ 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 [42]. 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 $\boldsymbol{Y} = \{Y_1, ..., Y_K\}$, 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 $\boldsymbol{y} = \{y_1, ..., y_K\}$ 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 $\boldsymbol{\pi} = \{\pi_1, ..., \pi_K\}$ as follows:

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

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

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

Where ω_1 and ω_2 are weighting factors to balance these two terms, and $\boldsymbol{\lambda}(\cdot)$ is the DSO's function to calculate the reward in terms of PAR. 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 $\boldsymbol{\lambda}$ 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. Moving to MARL, new relationships appear between agents in the same environment that compete or cooperate between them to maximize their rewards, as presented in Figure 3. 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 [44].

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

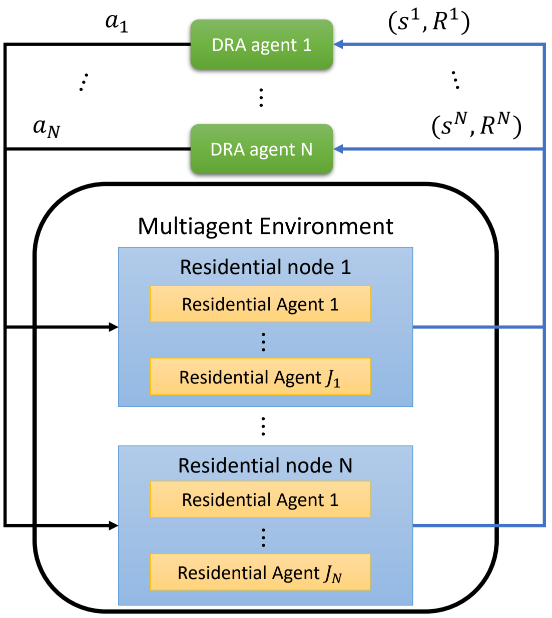


Figure 3: Multi-agent interaction with the same environment.

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 $R^{n,t}$.

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 [45]. 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 $\boldsymbol{\lambda}(\cdot)$ is based on the same proposed by [40], as follows:

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

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. All of them are determined during the negotiation between the DSO and the system operator. 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. Whit 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,

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

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. 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 [46], 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 [47]. 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 equations (3) and (5), the individual reward at state s^t for agent n can be finally stated as follows:

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

The Algorithm 1 represents the utilized IPPO technique.

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,

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

where the vector $\boldsymbol{u}^j = \{u_1^j, \cdots, u_K^j\}$ represents the consumption plan of the j^{th} 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

Algorithm 1: IPPO algorithm

```
For each DRA agent $n$:
DRA asks residential agents for their stipulated consumption plan under the initial constant price $\pi_0$, 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 $t = 0, 1, 2, ...$ do
    Define the action $a^{n,t} = \{M^n, \alpha^n\}$. (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 [48].
    end
end
```

and $x^j_{\max}$ are the lower and upper bounds for the allowed internal temperature according to users thermal preferences, respectively, and $u^j_{\max}$ 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^j_k, the outdoor temperature x^{out}_k, the heating power consumption $u^j_{h,k}$ and the matrix coefficients $\boldsymbol{w}^i$. According to [49, 50] this model can be expressed as:

$$\begin{aligned} x^j_{k+1} &= f(x^j_k, x^{\text{out}}_k, u^j_{h,k}, \boldsymbol{w}^j) \\ &= w^j_1 x^j_k + w^j_2 x^{\text{out}}_k + w^j_3 u^j_{h,k}. \end{aligned} \tag{11}$$

The first term in equation 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 $U(u^j_k)$ models the thermal user's thermal comfort and is determined by the set-point temperature x^j_{sp} and δ^j_k, the comfort weight factor representing the user's elasticity. δ^j_k explains how much users are willing to sacrifice their comfort to reduce the bill. According to [51], the residential thermal comfort function can be modeled through,

$$U(u^j_{h,k}) = -\delta^j_k (x^j_{\text{sp}} - x^j_k)^2, \tag{12}$$

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 [52]. 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 $t-1$ [53]. Thus, the dual optimization problem residential agents' cost function can be defined by (13).

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

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 equation (11). Furthermore, statistical information from a previous study conducted in [54] is utilized to randomly generate the set-point values x_{sp}^j from the set $\{20, 21, 22, 23\}$ in degree Celsius [C]. The different levels of users' thermal elasticity δ_k^j for the utility functions can be extracted from a log-normal distribution with the expectation, $\mathbb{E}(\delta_{\text{max}}) = 5$, and variance, $Var(\delta_{\text{max}}) = 1$. 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.

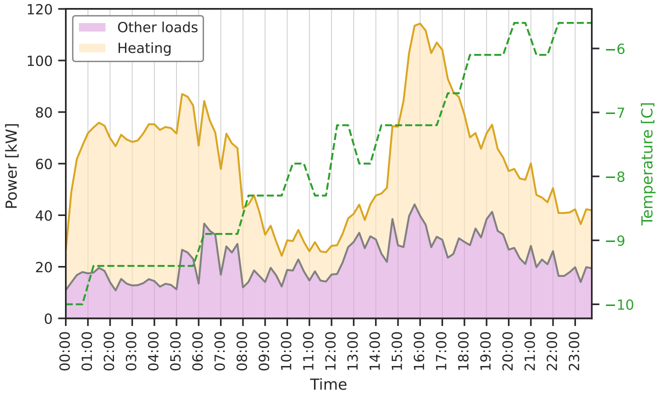


Figure 4: Aggregate energy demand when exposed to a winter outdoor temperature profile.

Figure 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.

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 [40]. 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, $\omega_2 = \sum_{k=1}^{N} \pi_k^n(a^{n,0})y_k^{n,0}$ 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.

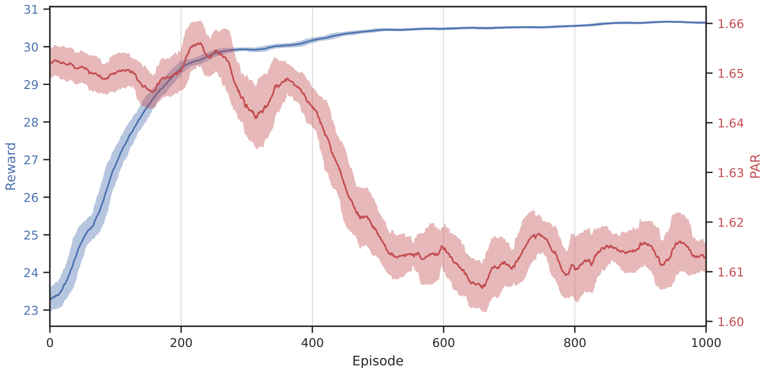


Figure 5: Analysis of the IPPO mechanism performance during the training process.

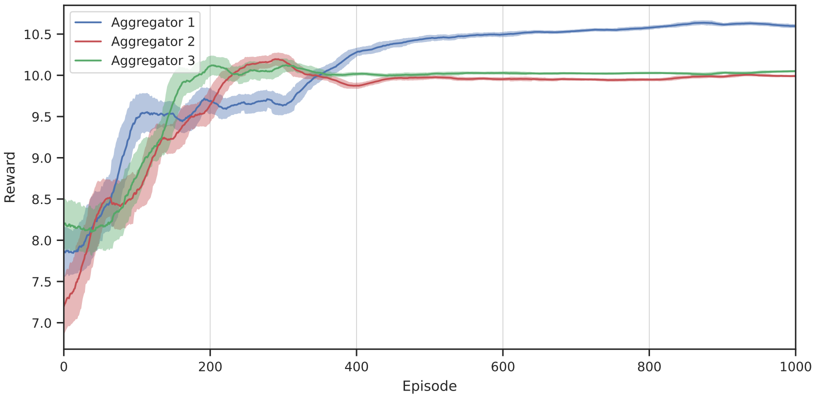


Figure 6: Analysis of the individuals DRAs training process.

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 the changes in consumption plans from one iteration to another are less than 10%. Figure 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, Figure 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 offering price discounts to the end-users allows the capitalization of the DSO's reward.

Figure 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. Figure 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.

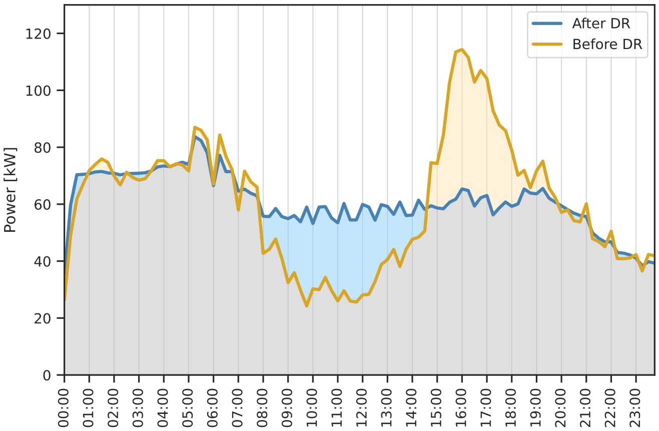
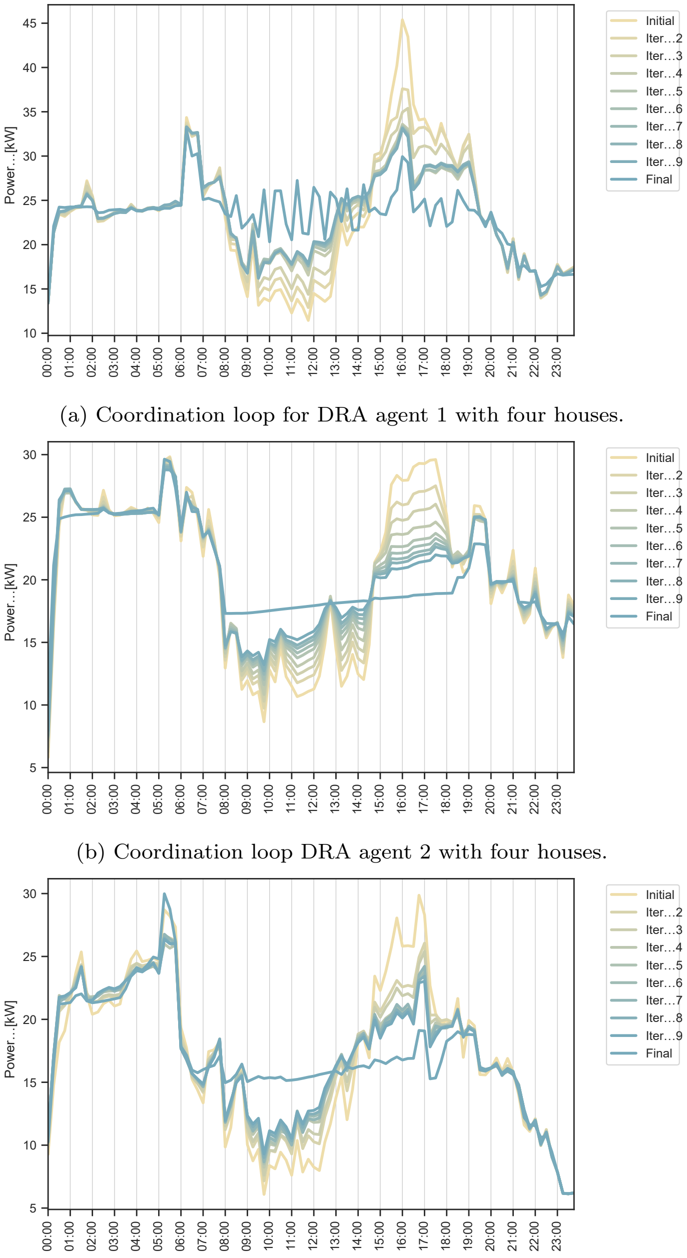


Figure 7: Peak reduction after learning process.

Shapley-Value-based reward-sharing mechanism. Finally, to analyze the importance of combining the IPPO algorithm with the SV-based reward-sharing mechanism, in Figure 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



(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.

Figure 8: DRAs' coordination loops after training.

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.

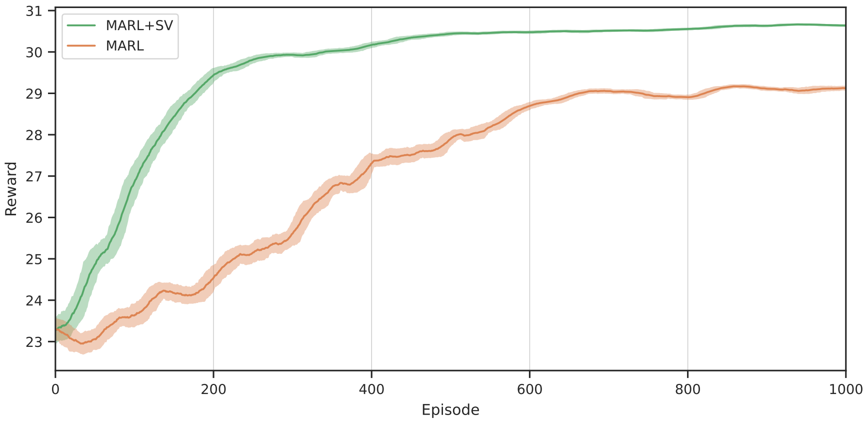


Figure 9: MARL performance with and without the SV-based reward-sharing mechanism.

Performance comparison. The proposed MARL-based mechanism is finally compared with a proximal decomposition approach proposed by the authors in [55]. 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 2 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.

		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%		

Table 2: Performance comparison between the proposed MARL-based mechanism and a proximal decomposition approach.

4. Conclusions

In this paper, a cooperative price-based demand response mechanism for a multi-aggregator system, based on multi-agent reinforcement learning (MARL) and a Shapley-Value-based reward sharing mechanism is proposed. This work utilized an IPPO-based MARL architecture for a set of demand response aggregator (DRA) agents to exploit the flexibility potential of residential customers. The DRAs establish dynamic pricing discounts in an iterative process, where DRAs communicate their price profiles and customers adapt in accordance with their consumption plan. In this win-win approach, the residential users leverage the flexibility of their controllable loads to reduce their bills, while the DRAs exploit this flexibility to reduce the system aggregated peak demand. This flexibility allows the DRAs to have access to the rewards offered by the DSO for peak reduction. The results presented demonstrate a significant PAR reduction in the 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%. Further, 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.

Acknowledgment

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

References

[1] S. Martinez, M. Vellei, J. Le Dréau, Demand-side flexibility in a residential district: What are the main sources of uncertainty?, Energy and Buildings 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 Transactions on Smart Grid 6 (4) (2015) 1874–1883. doi:10.1109/TSG.2014.2388357.

[3] A. R. Jordehi, Optimisation of demand response in electric power systems, a review, Renewable and sustainable energy reviews 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, Renewable and Sustainable Energy Reviews 77 (2017) 395–405. doi:https://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. doi:https://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 Transactions on Power Systems 36 (2) (2021) 979–990. doi: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. doi: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. doi: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 Informatics 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, Applied Energy 339 (2023) 121018. doi:https://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). doi: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 Transactions on Smart Grid 6 (6) (2015) 2837–2844. doi:10.1109/TSG.2015.2402194.

[15] L. S. Shapley, Notes on the N-Person Game &mdash; II: The Value of an N-Person Game, RAND Corporation, Santa Monica, CA, 1951. doi: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, Advances in neural information processing systems 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. doi: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, International Journal of Electrical Power & Energy Systems 130 (2021) 107004. doi:https://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, Applied Energy 187 (2017) 449–464. doi:https://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 Transactions on Smart Grid 13 (2) (2022) 1100–1112. doi: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. doi:10.1109/SmartGridComm.2011.6102350.

[23] L. Jia, L. Tong, Dynamic pricing and distributed energy management for demand response, IEEE Transactions on Smart Grid 7 (2) (2016) 1128–1136. doi: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, International Journal of Electrical Power & Energy Systems 118 (2020) 105764. doi:https://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 Transactions on Smart Grid 10 (1) (2019) 153–162. doi: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. doi:10.1109/EEEIC/ICPSEurope51590.2021.9584667.

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

[28] 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, Applied Energy 288 (2021) 116623. doi:https://doi.org/10.1016/j.apenergy.2021.116623.
URL https://www.sciencedirect.com/science/article/pii/S0306261921001586

[29] 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, Building and Environment 223 (2022) 109461. doi:https://doi.org/10.1016/j.buildenv.2022.109461.
URL https://www.sciencedirect.com/science/article/pii/S0360132322006928

[30] 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). doi:10.3390/en11092391.
URL https://www.mdpi.com/1996-1073/11/9/2391

[31] 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 Generation, Transmission & Distribution 14 (6) (2020) 1077–1090.

[32] 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, Applied Sciences 11 (11) (2021) 4948.

[33] J. Wang, W. Xu, Y. Gu, W. Song, T. C. Green, Multi-agent reinforcement learning for active voltage control on power distribution networks, Advances in Neural Information Processing Systems 34 (2021) 3271–3284.

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

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

[36] J. van Tilburg, L. C. Siebert, J. L. Cremer, Marl-idr: Multi-agent reinforcement learning for incentive-based residential demand response, arXiv preprint arXiv:2304.04086 (2023).

[37] 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.

[38] 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, International Journal of Electrical Power & Energy Systems 108 (2019) 72–85.

[39] F. Khavari, A. Badri, A. Zangeneh, Energy management in multi-microgrids via an aggregator to override point of common coupling congestion, IET generation, transmission & distribution 13 (5) (2019) 634–642.

[40] 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, Applied Energy 336 (2023) 120833. doi:https://doi.org/10.1016/j.apenergy.2023.120833. URL https://www.sciencedirect.com/science/article/pii/S0306261923001976

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

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

[43] 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) 100139doi:https://doi.org/10.1016/j.segy.2024.100139.

URL https://www.sciencedirect.com/science/article/pii/S2666955224000091

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

[45] 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.

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

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

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

[49] 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.

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

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

[52] G. Scutari, D. P. Palomar, F. Facchinei, J.-S. Pang, Monotone games for cognitive radio systems, Distributed decision making and control (2012) 83–112.

[53] 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. doi:10.1109/ACCESS.2022.3175839.

[54] 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.

[55] H. K. Nguyen, J. B. Song, Z. Han, Distributed demand side management with energy storage in smart grid, IEEE Transactions on Parallel and Distributed Systems 26 (12) (2015) 3346–3357. doi:10.1109/TPDS.2014.2372781.

Chapter 4 - Discussion and future opportunities

4.1 Introduction

This section extensively discusses the work carried out, the limitations and difficulties, and future opportunities based on the studies conducted for every targeted element. Accordingly, the following sections attempt to describe the statement of each problem as they remain open for future consideration. Each of these future opportunities will be categorized according to the energy market's perspective and the demand side's perspective in the smart grid context.

4.2 Work summary

- The analyses performed in this thesis have enabled the generation of price-based transactive policies for implementing DR programs considering a third-party entity called DRA. For this purpose, a Multi-agent system was built to reproduce the interaction between a set of houses and an aggregator. These interactions were constructed considering rational residential agents capable of reacting in an optimal manner to different economic incentives offered by the DRA. With this system development, the DRA exploits the interactions with the set of residential agents to characterize their price-responsive behavior. However, the main concern in this thesis is how to deal with the different sources of uncertainties related to DR program implementations and the consideration of the customers' privacy through the tariff generation process. For this reason, RL techniques were suggested as a valuable solution for dealing with uncertainty. The studies carried out during this work allowed us to determine that the DRA can use the DR as the only source of information needed as long as the rational response of the users is regularized in the multi-agent system. This regularization avoids the prisoner's dilemma in the system and guarantees the existence of a convergence point for the generation of pricing policies, considering the maximization of the DRA's profit.

- Once the necessary conditions have been identified to guarantee convergence in the interaction between the DRA and the residential agents, the next stage seeks to establish a price generator function. This function aims to consider the operational cost of the system and provide a voltage service for the DSO. This work proposes the use of a dynamic pricing function based on a sigmoid function in a coordination loop that allows offering a voltage capacity service to the DSO while taking into account the existing market regulations. This means that by using this function, it is possible to perform peak shaving in a controlled manner. At the same time, the function facilitates the implementation of the DR program during the smart grid transition process by considering the regulations of energy selling prices in the residential context. Finally, this approach reduces the calculation complexity by the DRA since this agent will have to define the optimized parameters for the pricing function instead of defining all the price values for each time step.
- The last stage of this work focuses on analyzing the definition of pricing policies considering a group of DRAs in a multi-aggregator system. In this process, the DSO offers a reward based on a global objective of the system, and a cooperative mechanism is established among the DRAs. This approach allows each aggregator to maximize its own profit while maximizing the overall system objective. For this case a MARL mechanism is proposed, which allows to take advantage of the conditions established in the first part of the work to guarantee the convergence of each group, as well as of the function proposed in the second stage to reduce the computational complexity of the method. In addition, a fair Shapley-value-based reward-sharing mechanism is proposed, which allows for

dealing with the non-stationarity problem of MARL algorithms and improves the performance of aggregators in defining pricing policies.

4.3 Limitations and difficulties

4.3.1 Residential Thermal model

One of the major limitations in the simulation process of the proposed scenario is centered on the thermal models of the houses. For our case, a linear model was proposed in order to reduce the algorithmic complexity of the mechanism and to facilitate the simulation process. However, the training times of the aggregators were quite high, which hindered the analytical studies. For this reason, it was not possible to conduct studies with more residential agents, especially for the multi-aggregator systems. In this specific case, the computational resources were often unable to carry out the calculation process to generate the consumption plans of the residential agents.

4.3.2 Day-to-day transition in the residential model

As presented in this thesis, generating aggregators' pricing policies is linked to the user's consumption response. This means that the models' response has an important impact on the convergence points of the proposed mechanisms. In this sense, one of the limitations of the proposed residential model is the lack of consideration of the day-to-day transition. Therefore, it is usual that each residential agent underestimates the consumption needed at the beginning and the end of each day. However, dealing with this problem is not so simple since it is necessary to define which mechanisms and information each residential agent should use in order to better estimate these transitions. Moreover, this consideration may further increase the computational complexity of the models.

4.3.3 Simulation environment

During the development of this thesis, different simulation tools were used to carry out the analyses shown. In particular, the simulations were performed using an object-based programming language called Python, which allowed the exploitation of existing resources, such as the Gym API [69] and the TensorForce library [70]. The former is a library developed by openAI that allows the creation of RL environments to develop and compare these algorithms by providing a standard API to communicate between RL agents and environments, and the latter is an open-source deep RL framework, provided with different algorithms that can be tested in the Gym API. Despite the good performance of these tools and the use of high-power hardware, the sequential simulation generated a high computational cost due to the computation time of the residential thermal models. This highlights the need for the development of strategies that allow the realization of distributed computation.

4.4 Recommendations

4.4.1 Energy market perspective

Since mitigating consumption peaks was our primary objective, using day-ahead markets was a more suitable option for the project, as these markets facilitate the generation of consumption plans for residential agents, avoiding the implementation of predictive models for pricing tariffs. However, as was presented in one of this thesis studies, customers can deviate from their stipulated consumption plans due to the uncertainties related to uncontrollable real-life events (receiving a visit from a friend, a sick user who must stay home, damage to the home, etc.). From this point of view, it is possible to consider the implementation of spot market strategies to alleviate the impact of these undesirable events, not only from the demand side but also from the supply side, which can face different operational grid problems. This means that the proposed day-ahead strategies can be combined with spot market mechanisms to enhance the performance of the DR programs throughout the execution of the day-based planned decisions.

Considering the end-users as market players, it is assumed in this work that users are interested in participating in the DR mechanism offered by the DRA. Thus, it would be an opportunity to conduct a study about the consumers' willingness to participate in the offered DR program, creating a competitive scenario between DRAs. This consideration will create a more realistic market interaction, where the aggregators' welfare will consider the monetary incentive to attract more customers' attention while considering the supply side's reward for exploiting the demand-side flexibility.

From the point of view of the DSO as a market player, in this work, we assumed that this player has a contract with the DRAs, offering monetary incentives for reducing the consumption peaks. In order to fairly distribute these rewards, a reward-sharing mechanism based on Shapley-value calculation is proposed. However, it is important to note that the combinatorial calculation on which it is based may become a new challenge for a large number of DRAs. It is for this reason that it is recommended to perform an analysis based on approximation mechanisms for the calculation of the marginal contributions of DRAs. On the other hand, in the proposed approach, it is assumed that the users participating in the program are connected to the same node of the power distribution network. However, once the topology of the power grid is considered, as well as customer positioning, it will be noticed that even if the PAR of the aggregated profile is improved, the grid stability can be compromised, as presented in [71]. Therefore, it is encouraged to address the relevance of integrating more in-depth power system analyses, taking into account aspects like power losses and congestion. The presented ideas evidence the need to perform further studies of determining DSO's reward allocation strategies in DR program applications considering the calculation complexity and the power grid architecture, which, consequently, must be examined under the integration of grid simulators.

4.4.2 Demand side perspective

From the demand side perspective, the residential agents developed in this work were able to react to different price signals. These agents can manage their controllable loads to determine the optimal strategies for decreasing their bills while respecting end-users' comfort preferences. The resulting decision was communicated to the DRAs, which was used as input for the determination of the following transactive policy, as well as for the characterization of their price-responsive behavior.

As the price-responsive behavior of the customers is linked to their level of flexibility, this one can be defined from two different aspects. The first one is the elasticity of the end-users, which refers to how much they are willing to sacrifice their comfort to reduce their bill. This is important as a fully inelastic user will never change its consumption pattern for any monetary incentive. It would be interesting to analyze the impact of these users on the performance of the implemented DR mechanism. The second aspect comprises the controllable load that the users can use to gain advantage of the economic incentive. In this work, the only controllable load considered was the space heating system, representing a big portion of the residential demand during winter in Quebec's context. However, other sources of flexibility can be considered to evaluate the performance of the described mechanisms. For instance, it would be possible to consider the management of the charging cycle of electric vehicles, the integration of thermal storage systems, or even the consideration of heat pumps instead of electric baseboards.

Regarding the controllable load considered in this work, a linear thermal model was utilized to describe the houses' thermal dynamics based on the users' thermal preferences. While this state-space model can reduce significantly the computational complexity of the modelization process, it only captures approximately the dynamics of the real system. It would be interesting to analyze the effect of considering more elaborate models that can consider the heat transfer through windows and doors and even the effect of the occupancy considering the human body as another source of heat.

Another important point is the role of the end-user in the system. In this work, the customers participating in the DR programs are consumers equipped with the intelligence to react rationally to economic stimulus. However, one of the important points in the smart grid concept is the emergence of prosumers, who can be consumers and producers of energy on the grid. These entities can add more dynamics to the power system as they can buy and sell energy if needed, as well as provide higher levels of flexibility to the power system. The definition of transaction policies in this context is a key opportunity for future studies. It provides a proper path to achieve one of the main objectives of the smart grid concept, which is the appropriate integration of distributed energy resources into the power grid.

Chapter 5 - Conclusions

5.1 Conclusions

One of the most crucial challenges for energy grid management is the reduction of the consumption peaks. For this reason, DR programs have appeared as a viable solution to mitigate this problem, giving the end-users a more active role in the system. This work has been focused on generating tools for the entity called DRA to define optimized transactive policies to perform a better operation of DR programs. In addition, some considerations have been added to this objective, which includes respect for users' privacy, the integration of the needs of the network, and the coordination of the DRAs during the transaction policy generation process. For the achievement of this objective, three important studies have been conducted as presented below:

1. In the first part of this essay, we have proposed a multi-agent system for the interactions between the DRAs and the residential agents. The conditions to guarantee convergence of a price-based DR program are established, with a DRA agent defining the price policies and a set of residential agents reacting in an optimal manner with the implementation of a regularized mechanism to ensure coordination. We developed a data-driven DRA for generating price policies based on discounts with minimal information exchange with end-user agents. The developed mechanism reduces the infrastructure needs for communication and maintains customer agents' privacy while avoiding dishonest reporting opportunities. An offline training phase has been proposed to improve the aggregator agent's performance in maximizing its reward. Finally, for this first approach, the time-consuming convergence of the RL was avoided, enabling the possibility of online implementation. For future works, incentive-based DR

programs can be explored from the aggregator's side, as well as the analysis of different sources of flexibility on the customer's side.

2. The second approach considers the power generation cost and the system's constraints in the transaction policy generation process. In this case, a price-based demand response strategy incorporating power capacity and market constraints to coordinate a set of residential agents is constructed. For this purpose, a price generation function is proposed to parameterize the price policy generation and enable a capacity service for the DSO while respecting existing market regulations. This work develops an RL-based DRA agent to exploit the electric heating system's potential for reducing the aggregated peak load of residential houses. A dataset of 11 houses with real-world data from Quebec's winter season is exploited to construct thermal models. The suggested approach effectively harnessed the flexibility of residential agents to optimize the profit of the DRA by fine-tuning the parameters of the price generation function throughout the coordination loop. The simulation outcomes revealed that the proposed DR strategy is able to address deviations in agents' consumption plans, leading to a substantial increase in DRA's profits. The comparative analysis highlighted the superiority of the proposed price-driven DR scheme and the employed PPO-based RL demand response aggregator. It converged to a solution that yielded a higher reward than the well-established A2C method. Furthermore, there is potential for exploring additional sources of flexibility and the integration of prosumers within the proposed DR program in future research.

3. Finally, a cooperative price-based demand response mechanism for a multi-aggregator system, based on multi-agent reinforcement learning and a Shapley-Value-based reward sharing mechanism, is proposed. This work utilized

an IPPO-based MARL architecture for a set of demand response aggregator (DRA) agents to exploit the flexibility potential of residential customers. Utilizing the proposed price generator function, the DRAs establish dynamic pricing discounts in an iterative process, where DRAs elaborate pricing structures, influencing customers to adjust their consumption plans accordingly. In this mutually beneficial strategy, residential users harness the adaptability of their manageable loads to decrease their bills. At the same time, DRAs capitalize on this adaptability to reduce the overall peak demand on the system. This adaptability empowers DRAs to access the incentives offered by the DSO for reducing peak demand. The findings reveal a substantial reduction in PAR for total power demand. Moreover, implementing the SV-based reward-sharing mechanism is demonstrated to enhance the solution's efficiency and reduce convergence time, dealing with the non-stationarity problem of MARL architectures. For future analysis, the suggested approach can be studied considering the positioning of users and aggregators in the power grid, for further evaluation of the stability of the system.

The above conclusions finalize the studies performed through this study related to the generation of price policies for implementing price-based DR programs. The future of this work is promising because it addresses a problem that is getting closer and closer to the existing power grids. The development paths of DR programs will depend on the production trends of the different technologies that will lead to the massive implementation of transactive energy.

References

[1] M. Behrangrad, “A review of demand side management business models in the electricity market,” *Renewable and Sustainable Energy Reviews*, vol. 47, pp. 270–283, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032115001860

[2] N. O’Connell, P. Pinson, H. Madsen, and M. O’Malley, “Benefits and challenges of electrical demand response: A critical review,” *Renewable and Sustainable Energy Reviews*, vol. 39, pp. 686–699, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032114005504

[3] [Online]. Available: https://www.hydroquebec.com/data/documents-donnees/pdf/sustainability-report-2021.pdf

[4] J.-T. Bernard, D. Bolduc, and N.-D. Yameogo, “A pseudo-panel data model of household electricity demand,” *Resource and Energy Economics*, vol. 33, no. 1, pp. 315–325, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0928765510000576

[5] [Online]. Available: https://www.hydroquebec.com/about/publications-reports/annual-report.html

[6] J. L. Mathieu, M. G. Vayá, and G. Andersson, “Uncertainty in the flexibility of aggregations of demand response resources,” in *IECON 2013 - 39th Annual Conference of the IEEE Industrial Electronics Society*, 2013, pp. 8052–8057.

[7] S. Chen and C.-C. Liu, “From demand response to transactive energy: state of the art,” *Journal of Modern Power Systems and Clean Energy*, vol. 5, no. 1, pp. 10–19, 2017.

[8] J. R. Vázquez-Canteli and Z. Nagy, “Reinforcement learning for demand response: A review of algorithms and modeling techniques,” *Applied Energy*, vol. 235, pp. 1072–1089, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261918317082

[9] B. Mohandes, M. S. E. Moursi, N. Hatziargyriou, and S. E. Khatib, “A review of power system flexibility with high penetration of renewables,” *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 3140–3155, 2019.

[10] K. T. Ponds, A. Arefi, A. Sayigh, and G. Ledwich, “Aggregator of demand response for renewable integration and customer engagement: Strengths, weaknesses, opportunities, and threats,” *Energies*, vol. 11, no. 9, 2018. [Online]. Available: https://www.mdpi.com/1996-1073/11/9/2391

[11] L. Gkatzikis, I. Koutsopoulos, and T. Salonidis, “The role of aggregators in smart grid demand response markets,” *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 7, pp. 1247–1257, 2013.

[12] A. Fraija, K. Agbossou, N. Henao, and S. Kelouwani, “Peak-to-average ratio analysis of a load aggregator for incentive-based demand response,” in *2020 IEEE 29th International Symposium on Industrial Electronics (ISIE)*, 2020, pp. 953–958.

[13] R. Deng, Z. Yang, M.-Y. Chow, and J. Chen, “A survey on demand response in smart grids: Mathematical models and approaches,” *IEEE Transactions on Industrial Informatics*, vol. 11, no. 3, pp. 570–582, 2015.

[14] D. Toquica, K. Agbossou, N. Henao, R. Malhamé, S. Kelouwani, and F. Amara, “Prevision and planning for residential agents in a transactive energy environment,” *Smart Energy*, vol. 2, p. 100019, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666955221000198

[15] P. Thorsnes, J. Williams, and R. Lawson, “Consumer responses to time varying prices for electricity,” *Energy Policy*, vol. 49, pp. 552–561, 2012, special Section: Fuel Poverty Comes of Age: Commemorating 21 Years of Research and Policy. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301421512005721

[16] R. D’hulst, W. Labeeuw, B. Beusen, S. Claessens, G. Deconinck, and K. Vanthournout, “Demand response flexibility and flexibility potential of residential smart appliances: Experiences from large pilot test in belgium,” *Applied Energy*, vol. 155, pp. 79–90, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261915007345

[17] Y. Gong, Y. Cai, Y. Guo, and Y. Fang, “A privacy-preserving scheme for incentive-based demand response in the smart grid,” *IEEE Transactions on Smart Grid*, vol. 7, no. 3, pp. 1304–1313, 2016.

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

[19] O. M. Longe, S. Rimer, K. Ouahada, and H. C. Ferreira, “Time programmable smart devices for peak demand reduction of smart homes in a microgrid,” in *2014 IEEE 6th International Conference on Adaptive Science & Technology (ICAST)*, 2014, pp. 1–6.

[20] I. K. Maharjan, “Demand side management: Load management, load profiling, load shifting, residential and industrial consumer, energy audit, reliability, urban, semi-urban and rural setting,” 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:166844513

[21] I. Lampropoulos, W. L. Kling, P. F. Ribeiro, and J. van den Berg, “History of demand side management and classification of demand response control schemes,” in *2013 IEEE Power & Energy Society General Meeting*, 2013, pp. 1–5.

[22] Q. Qdr, “Benefits of demand response in electricity markets and recommendations for achieving them,” *US Dept. Energy, Washington, DC, USA, Tech. Rep*, vol. 2006, p. 95, 2006.

[23] C. Eid, E. Koliou, M. Valles, J. Reneses, and R. Hakvoort, "Time-based pricing and electricity demand response: Existing barriers and next steps," *Utilities Policy*, vol. 40, pp. 15–25, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957178716300947

[24] Y. Chen, P. Xu, J. Gu, F. Schmidt, and W. Li, "Measures to improve energy demand flexibility in buildings for demand response (dr): A review," *Energy and Buildings*, vol. 177, pp. 125–139, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778818310387

[25] J. A. Momoh, *Smart grid: fundamentals of design and analysis*. John Wiley & Sons, 2012, vol. 63.

[26] F. C. Robert, G. S. Sisodia, and S. Gopalan, "A critical review on the utilization of storage and demand response for the implementation of renewable energy microgrids," *Sustainable Cities and Society*, vol. 40, pp. 735–745, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210670717307084

[27] X. Wang, A. Palazoglu, and N. H. El-Farra, "Operational optimization and demand response of hybrid renewable energy systems," *Applied Energy*, vol. 143, pp. 324–335, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261915000100

[28] B. Mohandes, M. S. El Moursi, N. D. Hatziargyriou, and S. El Khatib, "Incentive based demand response program for power system flexibility enhancement," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2212–2223, 2021.

[29] O. Erdinç, A. Taşcikaraoğlu, N. G. Paterakis, and J. P. S. Catalão, "Novel incentive mechanism for end-users enrolled in dlc-based demand response programs within stochastic planning context," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 2, pp. 1476–1487, 2019.

[30] H. Aalami, M. P. Moghaddam, and G. Yousefi, "Demand response modeling considering interruptible/curtailable loads and capacity market programs," *Applied Energy*, vol. 87, no. 1, pp. 243–250, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030626190900244X

[31] J. Saebi, H. Taheri, J. Mohammadi, and S. S. Nayer, "Demand bidding/buyback modeling and its impact on market clearing price," in *2010 IEEE International Energy Conference*, 2010, pp. 791–796.

[32] R. Tyagi and J. W. Black, "Emergency demand response for distribution system contingencies," in *IEEE PES T&D 2010*, 2010, pp. 1–4.

[33] V. Venizelou, N. Philippou, M. Hadjipanayi, G. Makrides, V. Efthymiou, and G. E. Georghiou, "Development of a novel time-of-use tariff algorithm for residential prosumer price-based demand side management," *Energy*, vol. 142, pp. 633–646, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544217317735

[34] S. Datchanamoorthy, S. Kumar, Y. Ozturk, and 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.

[35] X. Zhang, “Optimal scheduling of critical peak pricing considering wind commitment,” *IEEE Transactions on Sustainable Energy*, vol. 5, no. 2, pp. 637–645, 2014.

[36] R. Yu, W. Yang, and S. Rahardja, “A statistical demand-price model with its application in optimal real-time price,” *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 1734–1742, 2012.

[37] J. Ikäheimo, C. Evens, and S. Kärkkäinen, “Der aggregator business: the finnish case,” *Technical Research Centre of Finland (VTT): Espoo, Finland*, 2010.

[38] N. Mahmoudi, E. Heydarian-Forushani, M. Shafie-khah, T. K. Saha, M. H. Golshan, and P. Siano, “A bottom-up approach for demand response aggregators’ participation in electricity markets,” *Electric Power Systems Research*, vol. 143, pp. 121–129, 2017.

[39] “Study on the effective integration of distributed energy resources for providing flexibility to the electricity system,” Apr 2015. [Online]. Available: https://energy.ec.europa.eu/publications/study-effective-integration-distributed-energy-resources-providing-flexibility-electricity-system_en

[40] M. S. Bakare, A. Abdulkarim, M. Zeeshan, and A. N. Shuaibu, “A comprehensive overview on demand side energy management towards smart grids: challenges, solutions, and future direction,” *Energy Informatics*, vol. 6, no. 1, p. 4, 2023.

[41] F. Pallonetto, M. De Rosa, F. D’Ettorre, and D. P. Finn, “On the assessment and control optimisation of demand response programs in residential buildings,” *Renewable and Sustainable Energy Reviews*, vol. 127, p. 109861, 2020.

[42] B. Celik, R. Roche, D. Bouquain, and A. Miraoui, “Coordinated home energy management in community microgrids with energy sharing among smart homes,” in *ELECTRIMACS 2017*, 2017, pp. 1–6.

[43] ——, “Coordinated energy management using agents in neighborhood areas with res and storage,” in *2016 IEEE International Energy Conference (ENERGYCON)*. IEEE, 2016, pp. 1–6.

[44] ——, “Decentralized neighborhood energy management with coordinated smart home energy sharing,” *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6387–6397, 2017.

[45] H. K. Nguyen, J. B. Song, and Z. Han, “Distributed demand side management with energy storage in smart grid,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 12, pp. 3346–3357, 2014.

[46] A. Khalid, N. Javaid, A. Mateen, M. Ilahi, T. Saba, and A. Rehman, "Enhanced time-of-use electricity price rate using game theory," *Electronics*, vol. 8, no. 1, p. 48, 2019.

[47] H. Qiu, W. Gu, L. Wang, G. Pan, Y. Xu, and Z. Wu, "Trilayer stackelberg game approach for robustly power management in community grids," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4073–4083, 2020.

[48] K. Amasyali, Y. Chen, B. Telsang, M. Olama, and S. M. Djouadi, "Hierarchical model-free transactional control of building loads to support grid services," *IEEE Access*, vol. 8, pp. 219 367–219 377, 2020.

[49] C. Feng, Z. Li, M. Shahidehpour, F. Wen, and Q. Li, "Stackelberg game based transactive pricing for optimal demand response in power distribution systems," *International Journal of Electrical Power & Energy Systems*, vol. 118, p. 105764, 2020.

[50] G. Tsaousoglou, N. Efthymiopoulos, P. Makris, and E. V. Arigos, "Personalized real time pricing for efficient and fair demand response in energy cooperatives and highly competitive flexibility markets," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 1, pp. 151–162, 2019.

[51] H. T. Javed, M. O. Beg, H. Mujtaba, H. Majeed, and M. Asim, "Fairness in real-time energy pricing for smart grid using unsupervised learning," *The Computer Journal*, vol. 62, no. 3, pp. 414–429, 2019.

[52] N. Zhao, B. Wang, and M. Wang, "A model for multi-energy demand response with its application in optimal tou price," *Energies*, vol. 12, no. 6, p. 994, 2019.

[53] E. Dehnavi and H. Abdi, "Optimal pricing in time of use demand response by integrating with dynamic economic dispatch problem," *Energy*, vol. 109, pp. 1086–1094, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S036054421630576X

[54] R. Lu and S. H. Hong, "Incentive-based demand response for smart grid with reinforcement learning and deep neural network," *Applied energy*, vol. 236, pp. 937–949, 2019.

[55] B. J. Claessens, P. Vrancx, and F. Ruelens, "Convolutional neural networks for automatic state-time feature extraction in reinforcement learning applied to residential load control," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3259–3269, 2016.

[56] H.-M. Chung, S. Maharjan, Y. Zhang, and F. Eliassen, "Distributed deep reinforcement learning for intelligent load scheduling in residential smart grids," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2752–2763, 2020.

[57] R. Lu, S. H. Hong, and X. Zhang, "A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach," *Applied Energy*, vol. 220, pp. 220–230, 2018.

[58] D. Cao, W. Hu, J. Zhao, G. Zhang, B. Zhang, Z. Liu, Z. Chen, and F. Blaabjerg, "Reinforcement learning and its applications in modern power and energy systems: A review," *Journal of modern power systems and clean energy*, vol. 8, no. 6, pp. 1029–1042, 2020.

[59] M. A. Khan, A. M. Saleh, M. Waseem, and I. A. Sajjad, "Artificial intelligence enabled demand response: Prospects and challenges in smart grid environment," *IEEE Access*, vol. 11, pp. 1477–1505, 2023.

[60] X. Zhang, D. Biagioni, P. Graf, and 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.

[61] S. Zheng, Y. Sun, B. Li, B. Qi, K. Shi, Y. Li, and Y. Du, "Bargaining-based cooperative game among multi-aggregators with overlapping consumers in incentive-based demand response," *IET Generation, Transmission & Distribution*, vol. 14, no. 6, pp. 1077–1090, 2020.

[62] H. Taherian, M. R. Aghaebrahimi, L. Baringo, and S. R. Goldani, "Optimal dynamic pricing for an electricity retailer in the price-responsive environment of smart grid," *International Journal of Electrical Power & Energy Systems*, vol. 130, p. 107004, 2021.

[63] K. Aurangzeb, S. Aslam, S. M. Mohsin, and M. Alhussein, "A fair pricing mechanism in smart grids for low energy consumption users," *IEEE Access*, vol. 9, pp. 22 035–22 044, 2021.

[64] L. D. Collins and R. H. Middleton, "Distributed demand peak reduction with non-cooperative players and minimal communication," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 153–162, 2017.

[65] N. Henao, M. Fournier, and 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.

[66] Z. Li, Z. Tian, J. Wang, and W. M. Wang, "Extraction of affective responses from customer reviews: an opinion mining and machine learning approach," *International Journal of Computer Integrated Manufacturing*, vol. 33, no. 7, pp. 670–685, 2020.

[67] P. Atrazhev and P. Musilek, "It's all about reward: Contrasting joint rewards and individual reward in centralized learning decentralized execution algorithms," *Systems*, vol. 11, no. 4, p. 180, 2023.

[68] S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: a survey," *Artificial Intelligence Review*, pp. 1–49, 2022.

[69] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[70] A. Kuhnle, M. Schaarschmidt, and K. Fricke, "Tensorforce: a tensorflow library for applied reinforcement learning," 2017. [Online]. Available: https://tensorforce.readthedocs.io/en/latest/

[71] J. Dominguez, A. Parrado-Duque, O. D. Montoya, N. Henao, J. Campillo, and K. Agbossou, "Techno-economic feasibility of a trust and grid-aware coordination scheme," in *2023 IEEE Texas Power and Energy Conference (TPEC)*, 2023, pp. 1–5.

Appendix A - Résumé

A.1 Introduction

Actuellement, les différents défis environnementaux ont fait ressortir la nécessité de développer différentes stratégies pour surmonter les problèmes climatiques. En ce sens, les systèmes électriques ont rapidement évolué en raison du mode de fonctionnement des systèmes d'énergie électrique basés sur la production [1].En termes de limitations économiques et de considérations environnementales, la mise en œuvre traditionnelle de grands générateurs centralisés au sein d'un monopole est considérée comme non optimale et non soutenable [2].

Par exemple, selon le rapport 2021 d'Hydro-Québec, près de 40% de la consommation d'énergie de la province de Québec est demandée par le secteur résidentiel [4]. De plus, l'exposition à de longues périodes hivernales fait en sorte que la consommation des charges thermiques représente plus de 70% de la consommation résidentielle, tel qu'il est présenté dans la figure A.1 [3]. Pour cette raison, bien que les données montrent que la production d'énergie est suffisante pour répondre aux besoins des utilisateurs du réseau, il est possible que certains jours de la période hivernale, la demande de consommation dépasse la production pendant les heures de pointe [5].

Compte tenu de ce qui précède, l'idée de poursuivre avec un système de transaction à sens unique est devenue un concept obsolète. Cependant, l'intégration des technologies de l'information et de la communication comme l'internet des objets a permis de développer un nouveau concept appelé *resaue intelligent* (SG). Ce concept vise à utiliser ces technologies pour réaliser d'importantes économies d'énergie et améliorer la gestion du réseau électrique [8]. La prise en compte du rôle des utilisateurs dans la gestion de l'énergie, connue sous

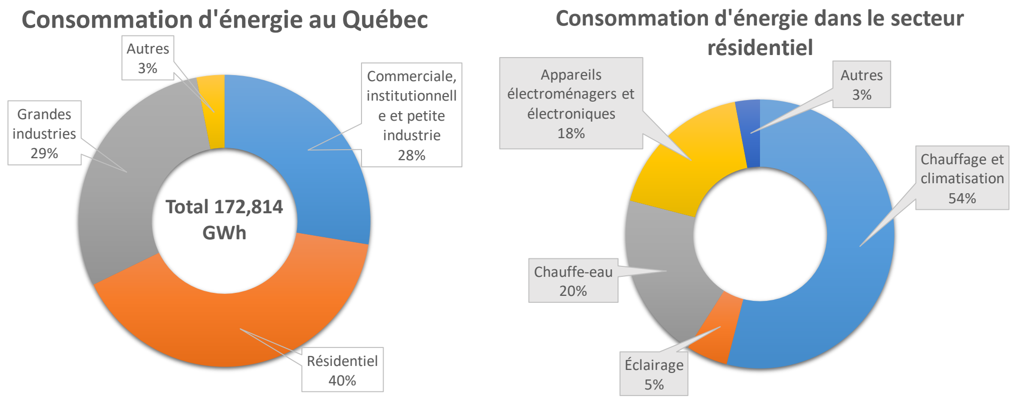


FIGURE A.1 – Consommation d'énergie au Québec en 2021, [3].

le nom de réponse à la demande (DR), est un élément important de la mise en œuvre du concept de SG. La DR est une modification de la consommation d'électricité par les clients par rapport à leur comportement de consommation normal. Il s'agit d'une réponse aux variations du prix de l'électricité ou à un paiement incitatif conçu par l'opérateur. L'idée est de réduire la consommation d'électricité lorsque les prix du marché de gros sont élevés ou lorsque la fiabilité du système est menacée [7].

Afin d'obtenir un comportement adéquat du réseau, une nouvelle entité appelée agrégateur DR (DRA) a été proposée. Sur le marché de l'électricité, un DRA est considéré comme un médiateur entre les acteurs du marché de l'électricité et les prosommateurs et consommateurs d'électricité. Ces agrégateurs proposent aux clients des contrats qui leur permettent de participer directement au marché de gros tout en fournissant aux opérateurs des services qui améliorent la fiabilité du réseau [12]. Ils exploitent les possibilités de flexibilité en gérant la consommation d'énergie résidentiel pendant les périodes critiques de la journée. Cela permet au DRA de capitaliser sur ces opportunités et participer sur les marchés de gros de l'électricité [10].

La participation de différents acteurs à l'échange d'informations et à l'optimisation distribuée permet de les considérer comme des agents déployés dans un environnement multi-agents, capables de négocier, de coordonner ou de coopérer en fonction des ressources qu'ils offrent ou dont ils ont besoin [13]. En conséquence, les agents peuvent atteindre l'équilibre du système en résolvant leurs problèmes d'optimisation individuels. Cette incroyable interaction, appelée énergie transactionnelle, est possible grâce à la communication bidirectionnelle du nouveau réseau. Cela crée un besoin supplémentaire de développement de nouvelles méthodologies innovantes pour surmonter les nouveaux défis liés à la nouvelle gestion de la consommation d'électricité [14]. Conformément au cadre présenté, la problématique de ce travail de recherche sera décrite ci-dessous.

A.2 Problématique de thèse

La nature humaine est l'un des principaux problèmes affectant la conception des marchés de DR. L'analyse du comportement des grands consommateurs d'énergie montre qu'ils réagissent de manière rationnelle en cherchant à maximiser leurs profits. Cependant, les petits consommateurs, tels que les utilisateurs résidentiels, ne font pas preuve de la même rationalité. En effet, les préférences de ces utilisateurs sont très différentes et, dans de nombreux cas, la minimisation de leur facture n'est pas forcément dans leur intérêt [2]. En outre, les auteurs de [15] ont mené une étude sur l'élasticité des prix des maisons, montrant qu'il n'existe pas de relation linéaire entre le changement de consommation et le changement de prix. Pour cette raison, la génération de politiques de prix optimisées reste un défi pour les programmes de DR afin d'exploiter le potentiel de flexibilité des utilisateurs [16].

Par conséquent, le DRA est chargée de surmonter ces problèmes afin de détecter de manière optimale les opportunités de flexibilité et d'en tirer parti. Il est donc nécessaire de développer des outils pour cet entité qui permettront définir des politiques transactionnelles optimisées afin d'atténuer les pics de consommation du secteur résidentiel [11]. Ces politiques devraient être adaptées grâce à la caractérisation des utilisateurs résidentiels qui interagiront avec le DRA. En conséquence, plusieurs défis se posent qui peuvent affecter la génération de ces politiques transactionnelles comme suit:

- Les problèmes d'incertitude dus à l'observabilité partielle des programmes de DR ont donné lieu à des considérations excessives concernant l'accès aux informations des utilisateurs pour l'élaboration d'une politique optimisée [17]. Ainsi, la génération de politiques de prix transactionnelles optimisées tout en évitant les impacts sur la vie privée des clients reste un défi.
- Le manque d'informations de la part des utilisateurs affecterait la convergence des méthodes mises en œuvre. En outre, la caractérisation du comportement du consommateur en réponse au prix devient un processus lent. Le défi consistant à assurer la convergence à un point proche de l'optimal tout en réduisant le temps de convergence est important pour garantir la viabilité des futures stratégies de prise de décision dans la génération de la politique de prix.
- Les besoins des fournisseurs peuvent être différents car ils peuvent être affectés non seulement par des aspects économiques mais aussi par les contraintes physiques du système. La grande majorité des études considèrent l'écrêtement des pointes comme une solution suffisante pour améliorer les performances du réseau. Cependant, ces stratégies prennent rarement en compte ces besoins réels, ce qui peut avoir un impact négatif sur l'optimisation économique des décisions du DRA.
- Enfin, il est nécessaire de prendre en compte l'augmentation de la complexité du système à mesure que le nombre de ces agrégateurs augmente. En plus, les

agrégateurs qui font partie de différentes entreprises doivent coopérer pour atteindre des objectifs globaux tout en maximisant leur rentabilité. Cela démontre la nécessité de développer des modèles coopératifs dans le processus de mise en œuvre des programmes de réduction de la consommation d'énergie.

A.3 Objectifs et contributions

En réponse aux problèmes présentés, ce projet a pour objectif de proposer des stratégies pour générer des politiques transactionnelles optimisées basées sur des agents intelligents, anticipant le comportement individuel et collectif des consommateurs d'énergie résidentiels. Les trois objectifs spécifiques suivants sont définis comme suit:

1. Générer des mécanismes permettant au DRA de définir des politiques transactionnelles optimisées en évitant les impacts sur la vie privée des clients et en augmentant l'intérêt des utilisateurs à participer au programme de DR.
2. Proposition de stratégies pour la génération de politiques transactionnelles optimisées basées sur les prix, intégrant les contraintes du marché et du système.
3. Développer un système multi-agents pour établir une méthode coopérative pour un ensemble de DRAs afin d'atteindre un objectif global du système tout en maximisant leurs propres profits.

La réalisation de ces objectifs se traduira par l'accomplissement des trois contributions principales suivantes :

1. La proposition d'une méthode permettant au DRA de générer des politiques transactionnelles optimisées pour les clients résidentiels en utilisant leur réponse aux politiques de prix et en traitant les incertitudes liées au manque d'informations domestiques.

2. La proposition de stratégies pour la génération de politiques transactionnelles optimisées basées sur les prix, intégrant les contraintes du marché et du système.
3. La proposition d'un système multi-agents coopératif, qui permettra à la gestion d'un ensemble de DRAs d'atteindre un objectif global du système tout en maximisant leurs profits.

En choisissant cette approche, de nouveaux défis apparaissent liés au manque d'informations disponibles pour la génération de politiques transactionnelles. C'est pourquoi, tout au long de cette thèse, nous cherchons à répondre à différents aspects, tels que :

- L'information minimale requise pour garantir la génération de politiques de tarification proche de l'optimal.
- Les conditions du système pour assurer la convergence ou l'équilibre de Nash.
- Le temps de convergence des algorithmes proposés et la proposition de méthodes pour les réduire (si nécessaire).
- La mise en œuvre de mécanismes de récompense équitables dans les approches coopératives de DRAs.

A.4 Méthodologie

Afin de traiter le problème évoqué, une approche basée sur des mécanismes pilotés par les données est proposée. Cette approche exploite l'interaction entre l'agent DRA et les agents résidentiels pour la génération de politiques transactionnelles. En ce sens, il s'agit d'une approche en trois étapes. La première consiste en une recherche bibliographique considérant le problème proposé pour comprendre et maîtriser les notions liées au domaine d'intérêt. Ceci sera fait en même temps que les modèles d'agents résidentiels sont développés afin qu'ils puissent interagir avec les agents DRA. Dans un deuxième temps, les limites et les difficultés des approches existantes seront analysées en tenant compte des exigences du problème traité. Cela permettra d'obtenir comme produit final une proposition qui fournit une solution appropriée au problème de recherche. Enfin, dans la troisième phase, la performance des propositions sera validée par des simulations et des mises en œuvre. Une illustration résumant la méthodologie suivie est présentée dans la figure A.2.

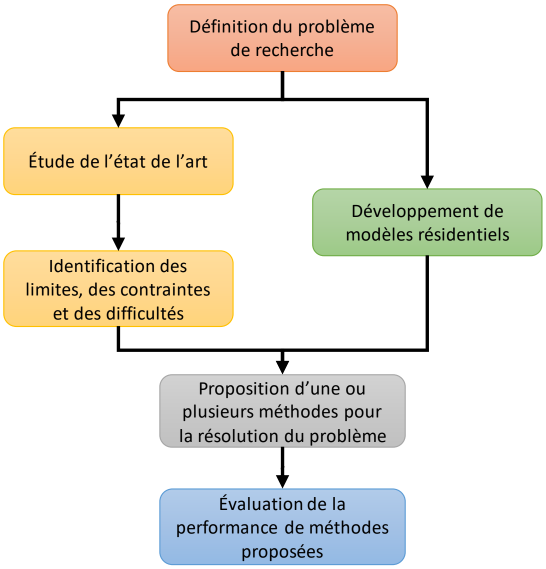

Définition du problème de recherche
Étude de l'état de l'art
Développement de modèles résidentiels
Identification des limites, des contraintes et des difficultés
Proposition d'une ou plusieurs méthodes pour la résolution du problème
Évaluation de la performance de méthodes proposées

FIGURE A.2 – Méthodologie de recherche suivi.

Ainsi, pour atteindre les objectifs de ce projet de thèse, en suivant l'approche en trois étapes proposée, les activités suivantes seront réalisées : Tout d'abord, nous développerons des modèles pour les charges contrôlables que les agents résidentiels pourront utiliser comme source de flexibilité. Ensuite, un modèle comportemental sera construit pour ces agents, et un mécanisme de contrôle sera défini en utilisant les charges contrôlables pour obtenir un comportement réactif aux stimuli (politiques transactionnelles, météo, etc.). Enfin, nous analyserons comment le DRA peut définir des politiques transactionnelles optimisées en fonction de la réponse de la consommation d'un ensemble donné d'agents résidentiels. Le DRA caractérisera le comportement réactif des utilisateurs finaux en apprenant de leurs interactions.

A.4.1 Hypothèses de recherche

Les hypothèses de recherche suivantes seront prises en compte lors de l'élaboration des mécanismes proposés pour atteindre les objectifs du problème de recherche décrit:

- On suppose que le mécanisme de coordination est suffisamment rapide pour ne pas interférer avec le temps d'exécution du programme de DR. Cela signifie que la politique transactionnelle sera définie à temps pour que les clients puissent ajuster leurs plans de consommation pour la prochaine période financière.
- Dans le système de communication, on suppose que l'agent agrégateur envoie équitablement la même information à tous les agents résidentiels. En outre, le marché fonctionne de telle manière qu'il ne permet pas l'échange d'informations entre ces derniers, dans le but d'abuser de ces informations supplémentaires pour augmenter leur propre profit.
- Enfin, il est important de souligner l'hypothèse adoptée de la rationalité économique du consommateur, qui est à la base de la mise en œuvre des programmes de DR. En outre, il est supposé que les agents résidentiels sont capables de réagir de manière optimale, conformément à la rationalité du consommateur.

A.5 Description de résultats publiés

Les résultats de la méthodologie proposée pour atteindre les objectifs de ce projet de recherche ont été séparés en trois publications. Tout d'abord, un mécanisme de tarification est développé pour évaluer l'architecture du système multi-agents afin de garantir la convergence, puis une fonction de génération de prix est proposée pour paramétrer la génération de la politique de prix en tenant compte des contraintes du marché et de la fourniture, et une architecture multi-agrégateurs est établie en utilisant l'architecture multi-agents et la fonction de génération de prix proposées. Le statut de publication des articles présentés ci-dessous est le suivant :

1. Le premier article a été publié dans IEEE Access le 17 mai 2022.
2. Le deuxième a été publié dans Smart Energy d'Elsevier le 27 mars 2024.
3. Enfin, le troisième manuscrit a été soumis à Sustainable Energy, Grids and Networks d'Elsevier le 9 avril 2024, et est actuellement en cours de révision.

A.5.1 Architecture multi-agents pour la génération de politiques transactionnelles

Contexte:

Dans cette première partie, nous examinons le problème du développement d'un système multi-agents pour l'interaction entre les agents résidentiels et un agent du DRA en tant qu'acteurs du marché. L'interaction entre ces entités permettra au DRA de rechercher l'optimisation au cours du processus de génération de la politique de prix. Il faut donc construire des agents résidentiels automatisés pour garantir des réponses rationnelles aux actions des agents du DRA. Dans le cas de l'agent du DRA, la seule source d'information pour lui sera le DR afin de respecter la vie privée des clients.

Le programme DR mis en place par le DRA sera une stratégie de tarification de l'électricité basée sur des rabais. Cet agent offrira des rabais à différentes heures de la journée afin d'encourager les utilisateurs à modifier leurs habitudes de consommation. Le processus de prise de décision est réalisé en appliquant une technique apprentissage par renforcement (RL) profonde pour traiter les différentes sources d'incertitude dues au manque d'informations du côté de la demande. l'agent DRA essaiera de maximiser le facteur de charge (inverse du PAR) tout en minimisant la réduction de son profit pendant le processus de rabais.

Pour l'agent résidentiel, un modèle thermique est construit pour les systèmes de chauffage. Ce modèle permettra d'utiliser le chauffage comme source de flexibilité pour les utilisateurs. L'agent résidentielle déterminera son plan de consommation en réponse à la politique de prix en résolvant un problème d'optimisation. L'objectif de l'agent résidentiel est de minimiser sa facture tout en maintenant le confort thermique des clients. Enfin, un mécanisme de régularisation de la réponse de l'agent résidentiel est appliqué pour garantir la convergence du système multi-agents.

Méthodologie:

L'environnement RL est composé d'un ensemble de vingt maisons résidentielles. Un travail antérieur est utiliser pour déterminer les préférences thermiques des clients [65]. En exploitant les données historiques collectées, les paramètres du modèle thermique sont estimés à l'aide d'un mécanisme de régression rigide [66]. Enfin, un processus de génération de données statistiques est utilisé pour les charges non contrôlables.

Une fois que les agents résidentiels sont prêts à répondre aux signaux de prix, une journée historique est sélectionnée pour l'entraînement hors ligne du DRA basé sur le RL. Les analyses suivantes ont été effectuées pour garantir la meilleure performance du programme de DR basé sur les prix proposé :

1. Comparaison des performances de différentes techniques Deep RL pour sélectionner le PPO comme mécanisme cible pour la mise en œuvre du programme de DR.

2. Analyse de la convergence du système multi-agents basée sur la sélection du paramètre de régularisation pour les agents résidentiels.
3. Une étude comparative entre l'approche proposée basée sur RL et un mécanisme de décomposition proximale tiré de la littérature en termes d'amélioration du facteur de charge et de profit du DRA.

Après avoir effectué ces évaluations hors ligne, les performances du processus décisionnel du DRA ont été évaluées en ligne pendant plusieurs jours consécutifs.

Résultats:

L'approche présentée a développé un DRA basé sur des données pour générer des tarifs horaires ToU proches de l'optimum. Cette conception permet de réduire les besoins en infrastructures de communication et de préserver la vie privée des agents de clientèle dans le cadre d'interactions fiables.

En termes de mise en œuvre, cette étude a recommandé un algorithme RL pour construire un système DR prometteur. En outre, l'approche proposée fournit une stratégie de phase d'entraînement hors ligne pour traiter la convergence des techniques RL. Cette proposition a permis de réduire le temps de convergence de plus de 1000 jours à moins de 20 jours.

Les résultats obtenus ont été comparés à deux techniques RL courantes. Le mécanisme RL proposé montre des performances supérieures en terme de taux de convergence. Un schéma de coordination est également comparé, où le DRA basé sur le RL peut obtenir une réduction plus faible de son profit, bien que le tarif optimisé soit basé sur des rabais. D'autre part, une réduction plus importante des revenus de la méthode de coordination montre que le sacrifice monétaire dans un programme de DR peut être élevé s'il n'est pas contrôlé. Ces résultats soulignent l'efficacité du mécanisme proposé.

A.5.2 Fonction générateur de prix

Contexte:

Désormais, un mécanisme de tarification dynamique sera développé pour améliorer l'exploitation des potentiels de flexibilité du côté de la demande, puisqu'une politique de prix plus détaillée sera établie sur le marché day-ahead. L'idée est de déterminer une politique de prix tenant compte des contraintes de la fourniture, en ce qui concerne les limitations de capacité, tandis que le DRA atteindra ses objectifs en tenant compte des contraintes de prix du marché dans le processus de génération de la politique.

L'objectif principal de cette phase est de dériver un mécanisme de tarification dynamique pour offrir un service de limitation de capacité au DSO. Les services de capacité dans un contexte de tarification sont généralement offerts par le biais de mécanismes d'appel d'offres, ce qui entraîne des coûts de calcul élevés et une dépendance excessive à l'égard des informations sur les clients. En outre, la littérature ne prend pas en compte les limites de prix du marché existant, ce qui pourrait créer un impact significatif sur les processus d'optimisation des approches existantes.

Pour cela, une fonction générateur de prix dynamique est proposée, prenant en compte les contraintes de la fourniture et du marché dans un scénario de théorie des jeux mettant en œuvre une boucle de coordination. Grâce à cette fonction, le DRA sera en mesure de maintenir les besoins en capacité du DSO. En réponse, le DSO paiera une incitation au DRA. Le DRA essaiera de maximiser son utilité, en tenant compte le bénéfice de la vente de l'électricité aux clients et de l'incitation du DSO. Ensuite, le DRA utilisera une technique RL pour définir les paramètres de cette fonction tout à long de la boucle de coordination, en tenant compte non seulement des besoins du DSO, mais aussi de la déviation possible du profil de consommation des clients par rapport à leurs plans de consommation stipulés.

Méthodologie:

Le DSO communique la limite de capacité désiré et offre une incitation au DRA sur la base d'une fonction de coût de production d'électricité quadratique. Cette incitation est déterminée sur la base de la réduction du coût de production d'électricité due à la réduction de l'écrêtement des pointes. Ensuite, le DRA utilise la fonction de génération de prix proposée dans une boucle de coordination itérative au début de la journée. Le DRA communique d'abord un profil de prix constant et attend la réponse des agents résidentiels. Avec ce profil agrégé, le DRA calcule la politique de prix suivante en utilisant la fonction de génération de prix jusqu'à ce qu'un accord soit atteint.

La combinaison de cette fonction de génération de prix dans le système multi-agents développé, avec la boucle de coordination, crée une tendance qui fait que la pointe de consommation maximale se situe dans un voisinage centré sur la limite de capacité établie par le DSO avec un rayon qui dépend du niveau d'élasticité des utilisateurs. Avec ce comportement, les étapes suivantes ont été suivies:

1. L'évaluation de la fonction génératrice de prix proposée en ce qui concerne la réduction de l'écrêtement des pointes.
2. Comparaison des performances de la fonction génératrice de prix proposée avec une fonction linéaire par morceaux en termes de dépassement de la limite de capacité et d'exploitation de la flexibilité.
3. Comparaison de la technique RL entre le mécanisme PPO sélectionné et la méthode A2C populaire.

4. Évaluation de la méthode pour gérer les écarts de consommation des clients tout en maximisant son utilité.

Résultats:

Ce travail fournit une fonction de génération de prix pour paramétrer le processus de génération de la politique de tarification dynamique. Cette fonction démontre une exploitation plus élevée et contrôlé du potentiel de flexibilité de la demande, permettant l'offre d'un service de capacité pour le DSO. En outre, elle prend en compte les régulations existantes du marché dans la génération des taux de tarification dynamique, garantissant la mise en œuvre de ce mécanisme dans des contextes réalistes du marché de l'énergie.

Des simulations sont effectuées pour évaluer les performances de la stratégie proposée, basée sur le RL. La mise en œuvre de la méthode proposé démontre que le DRA peut gérer les écarts des agents par rapport à leurs plans de consommation, tout en améliorant l'utilisation de la fonction génératrice de prix, puisque les bénéfices du DRA ont augmenté de plus de 30%. En ce qui concerne la sélection du mécanisme RL, la méthode PPO adoptée a convergé vers une solution qui fournit des récompenses plus élevées pour le DRA.

A.5.3 Système multiagrégateur

A.5.4 Contexte:

Avec l'environnement multi-agents défini pour les interactions entre un DRA et un ensemble de maisons et la définition d'une fonction génératrice de prix pour paramétrer le mécanisme de prix dynamique, il est maintenant temps d'évaluer cette génération de politique du point de vue du système. Les solutions individuelles des différents DRA ne garantissent pas l'obtention d'une bonne solution optimisée. Ce travail approfondit la question de la tarification dynamique avec plusieurs DRA sur la base de l'apprentissage par renforcement multi-agents (MARL), étant donné que chaque agrégateur continuera à interagir avec son propre ensemble de maisons.

Pour que les DRAs se coordonnent, il est nécessaire de mettre en œuvre un mécanisme d'attribution de récompenses équitables, en fonction de leurs contributions à l'objectif du système. C'est là que la valeur de Shapley (VS), un concept issu de la théorie des jeux coopératifs, entre en scène. Chaque DRA recevra une récompense du DSO en fonction de sa contribution à l'objectif global par le biais du calcul du SV. L'intégration de cette stratégie fournit un cadre équitable pour la répartition des avantages liés à la coopération entre les agents DRA. En outre, ces travaux démontrent également que l'évaluation des récompenses en fonction de leur impact marginal sur le système global accélère les performances de l'architecture MARL et aide à traité le problème de non-stationnarité de ces algorithmes [67].

Méthodologie:

Pour évaluer le mécanisme, il est nécessaire de construire l'environnement MARL en tenant compte du système multi-agrégateur. Sur la base du développement précédent de cette thèse, un ensemble de DRA est établi, chacun interagissant avec un ensemble de maisons avec une cardinalité différente. En outre, chaque client aura ses préférences en matière de confort, ce qui influencera sa réponse aux signaux transactionnels générés par chaque DRA. Par conséquent, chaque agrégateur doit apprendre ses stratégies dans une architecture MARL décentralisée (DTDE) [68].

Chaque agent DRA utilisera la fonction de prix pour offrir des réductions afin d'inciter les clients à modifier leurs habitudes de consommation. Les résultats de cette interaction établiront le tarif dynamique et le plan de consommation des clients pour les 24 heures suivantes. À la fin de la journée, le DSO déterminera la contribution marginale à la réalisation de l'objectif global du système pour chaque DRA, et c'est sur cette base que les récompenses des agrégateurs seront définies.

Résultats:

Ce travail propose un mécanisme coopératif de réponse à la demande pour un système multi-agrégateur basé sur MARL et un mécanisme de partage des récompenses sur la base de SV. Comme les DRA établissent des réductions de prix dynamiques dans un processus itératif, les clients peuvent adapter leurs profils de consommation pour bénéficier de ces réductions. Cette stratégie crée une approche gagnant-gagnant, car les utilisateurs résidentiels peuvent exploiter la flexibilité de leurs charges contrôlables pour réduire leurs factures, tandis que les DRA peuvent offrir cette flexibilité au DSO pour réduire la demande de pointe agrégée du système. Au moyen de la stratégie MARL, les DRA font un compromis entre la réduction de leur profit en offrant des réductions aux maisons pour exploiter leur potentiel de flexibilité et la récompense que le DSO leur offre pour l'effort réalisé. Les résultats présentés démontrent une réduction significative de la point totale de consommation. En outre, l'importance de la mise en œuvre du mécanisme de partage des récompenses basé sur SV est démontrée en termes d'amélioration de l'optimisation de la solution et de réduction du temps de convergence de la méthode.

A.6 Conclusions

Ce travail s'est concentré sur la création d'outils permettant à l'entité appelée DRA de définir des politiques transactionnelles optimisées afin d'améliorer le fonctionnement des programmes de DR. En outre, certaines considérations ont été ajoutées à cet objectif, notamment le respect de la vie privée des utilisateurs, l'intégration des besoins du réseau et la coordination des DRAs au cours du processus de génération de la politique transactionnelle. Pour atteindre cet objectif, trois études de recherche ont été abordées à savoir i) Le système multi-agents pour les interactions entre les DRAs et les agents résidentiels, ii) les contraintes du système et du marché dans le processus de génération de la politique transactionnelle et iii) les mécanismes de coopération pour un système à agrégateurs multiples. Les approches proposées ont été décrites dans trois publications sous forme d'articles scientifiques.