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

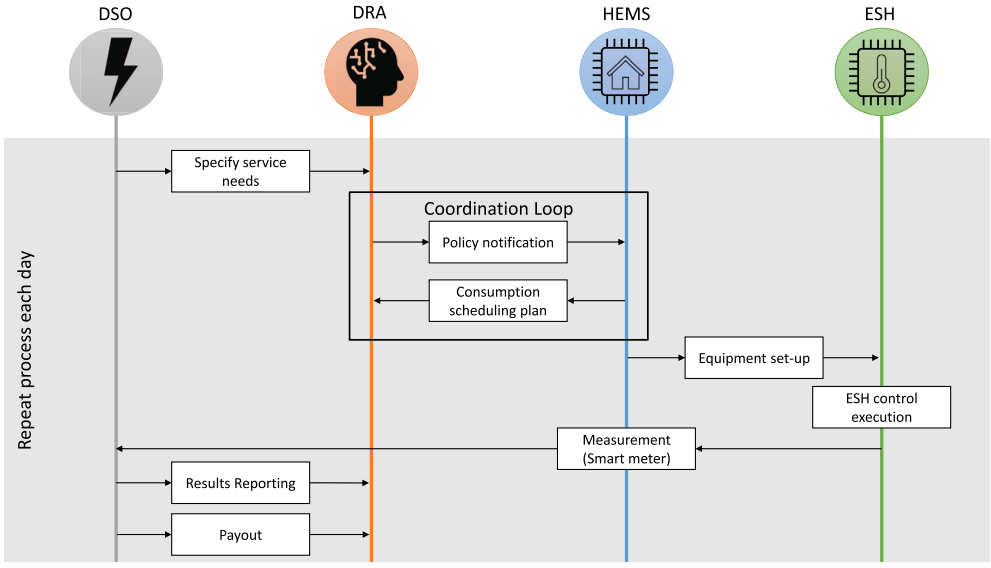


Fig. 1. Automatic price-based DR sequence.

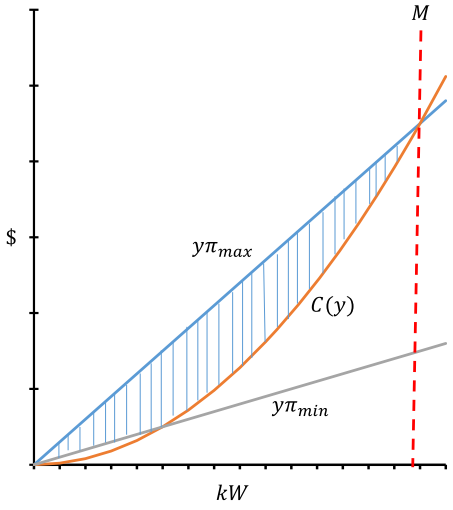


Fig. 2. Market and power constraints in terms of power generation cost function.

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,

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

where y_k represents the aggregate consumption at time stamp $k \in \{1, \ldots, N\}$. 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.

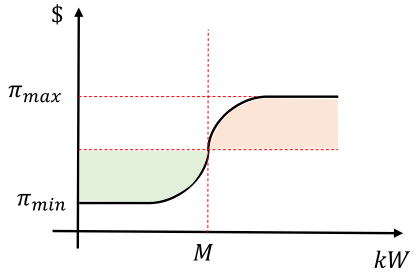


Fig. 3. Proposed price generator function.

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

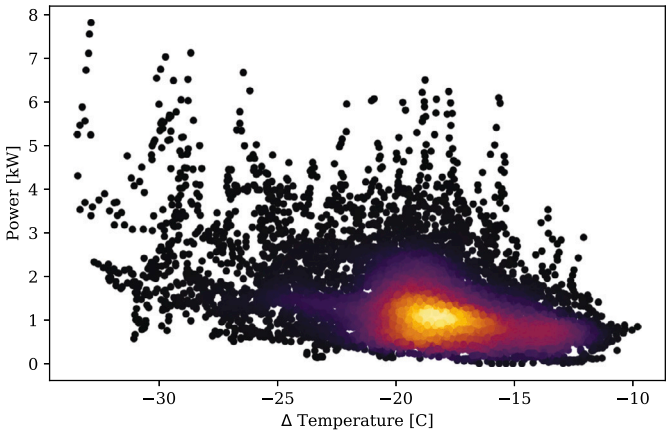


Fig. 4. Distribution of the ESH power consumption and the outdoor temperature for one house in Trois-Rivieres, Quebec.

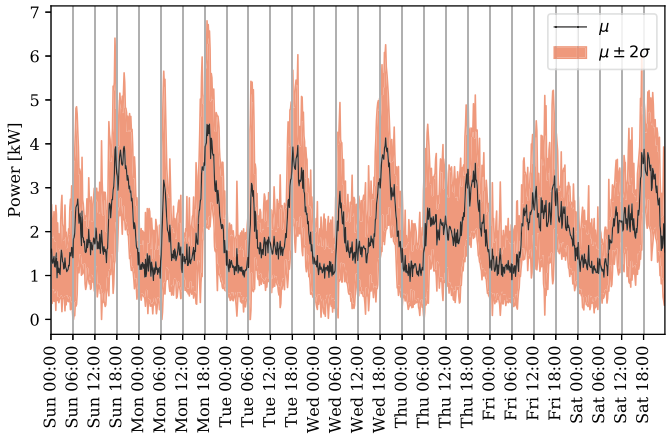


Fig. 5. Average weekly power profile from 8 real houses (space heating load is not included).

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.

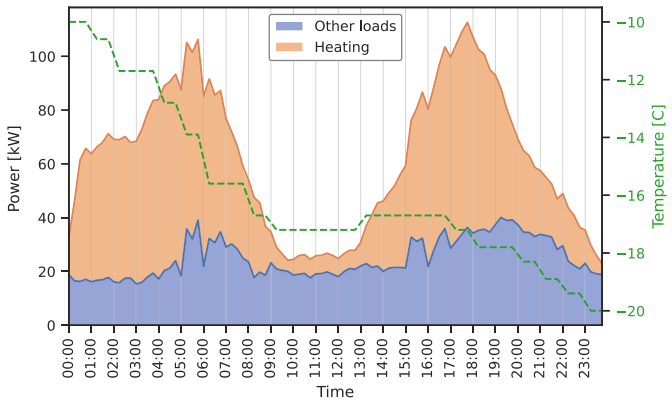


Fig. 6. One-day aggregated power demand without DR.

4.1. The scenario without DR

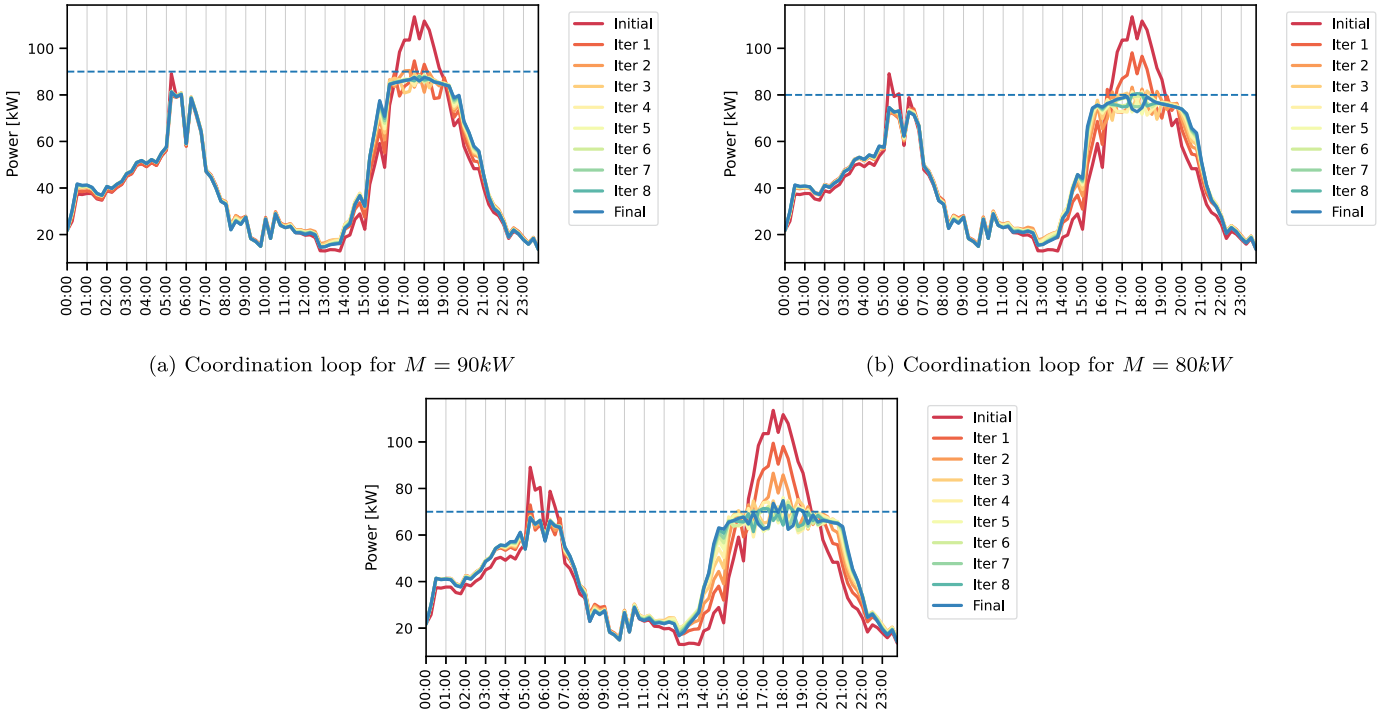
Fig. 6 shows the aggregated consumption profile of a set of 11 simulated buildings during a cold day. The consumption behavior in the figure demonstrates that the models developed are in accordance with the expected power consumption pattern in Quebec's residential sector. Each residential agent performs 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

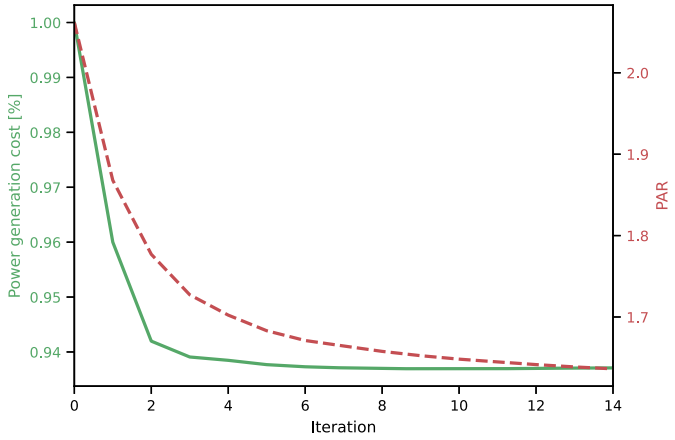


Fig. 8. Power generation cost percentage and PAR curves during coordination loop.

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.

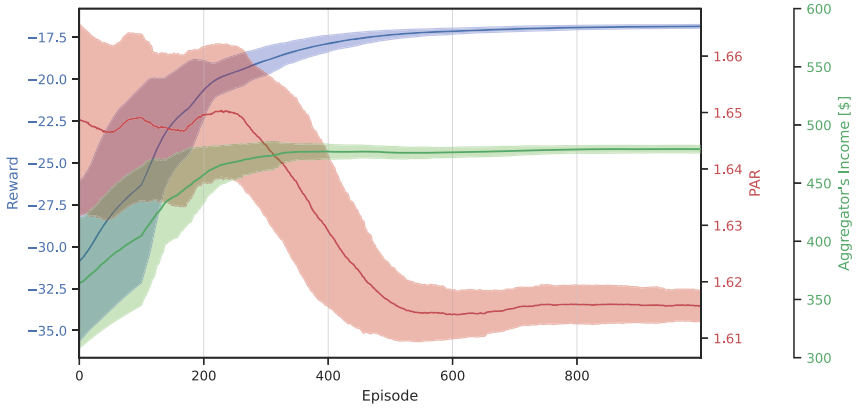


Fig. 9. Analysis of DRA agent performance during the learning process.

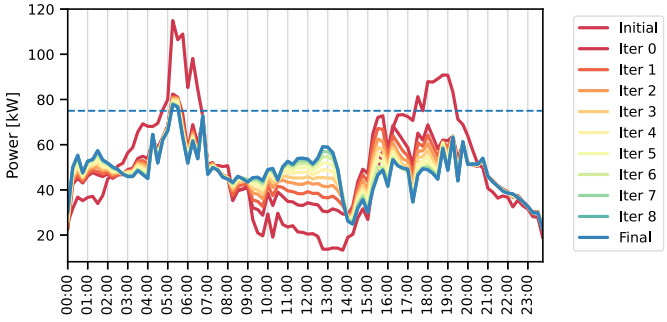


Fig. 10. Coordination loop after RL learning process with $M = 75kW$.

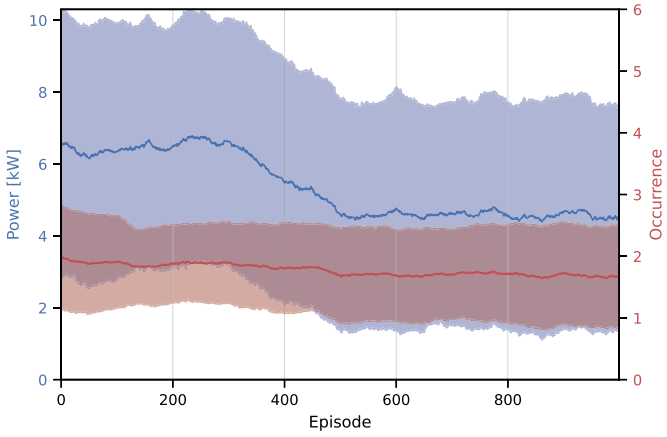


Fig. 11. Analysis of average capacity constraint overruns.

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).

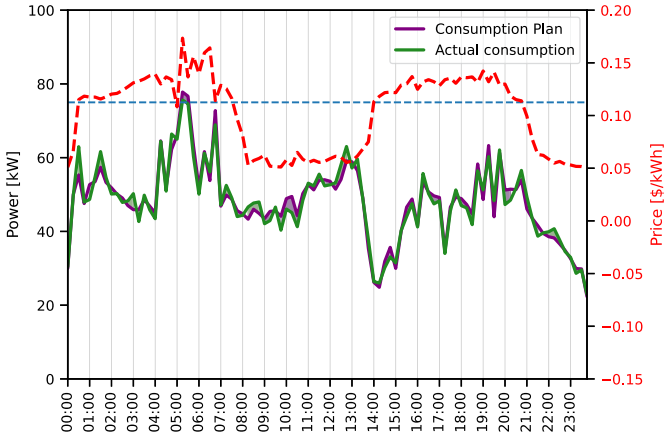
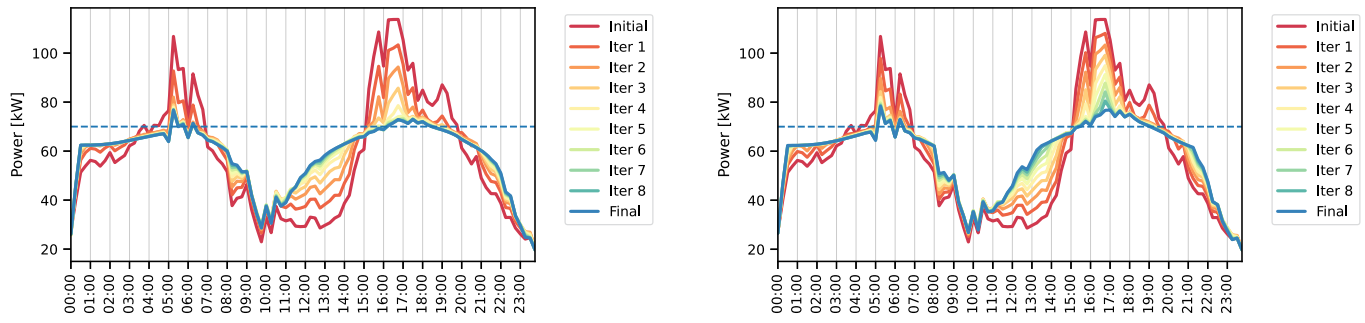


Fig. 12. Difference between actual aggregate consumption and the consumption plan of the agreement under the established price profile.

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.

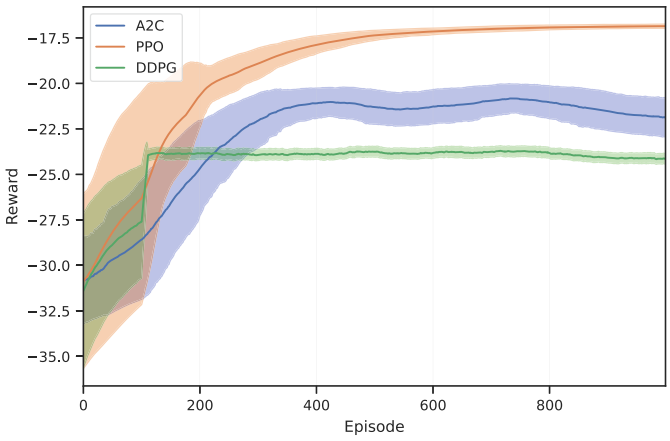


Fig. 14. Comparative performance of our PPO mechanism with A2C and DDPG.

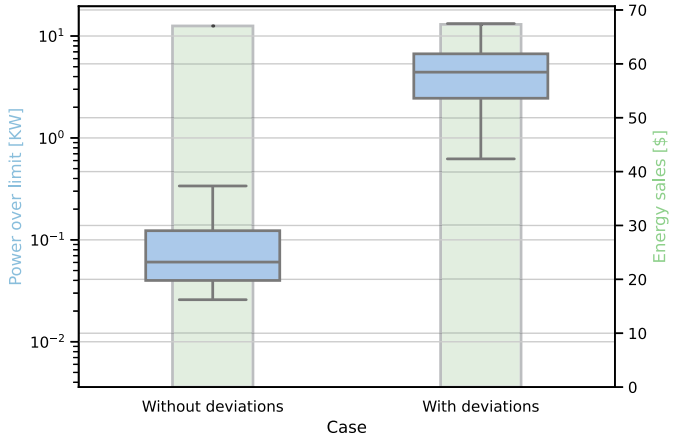


Fig. 15. Performance analysis related to the consideration of users' deviations from consumption plans.

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.