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.

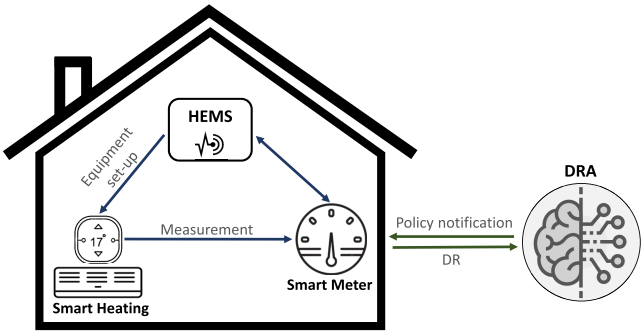


FIGURE 1. Automatic energy management under the price-based DR program.

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 θ_0, clipping threshold ϵ, 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.

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

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 $\lambda_t^h \leq \xi$. 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,

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

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,

$$\underset{u^i=\{u_k^i\}_{k=1}^T}{\text{Minimize}} \sum_{k=1}^{N} \delta_k(x_{\text{sp}}^i - x_k^i)^2 + \lambda_k u_k^i + \tau(u_{t,k}^i - u_{t-1,k}^i)^2$$

subject to Eqnarray(8)

$$\begin{aligned} x_k^i &\in [x_{\min}^i, x_{\max}^i] \\ u_k^{i,Th} &\in [0, u_{\max}^{i,Th}] \\ u_k^i &= u_k^{i,Th} + u_k^{i,NC} \end{aligned} \quad (11)$$

where the parameters $x_{\min}^i$ and $x_{\max}^i$ are the lower and upper bounds of the allowed internal temperature, respectively, and $u_{\max}^{i,Th}$ 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.

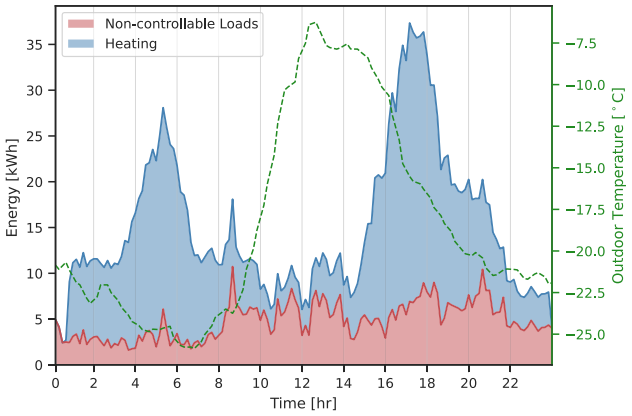


FIGURE 2. Aggregated energy consumption behavior of twenty residential agents in correlation with outside temperature on a typical winter day.

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

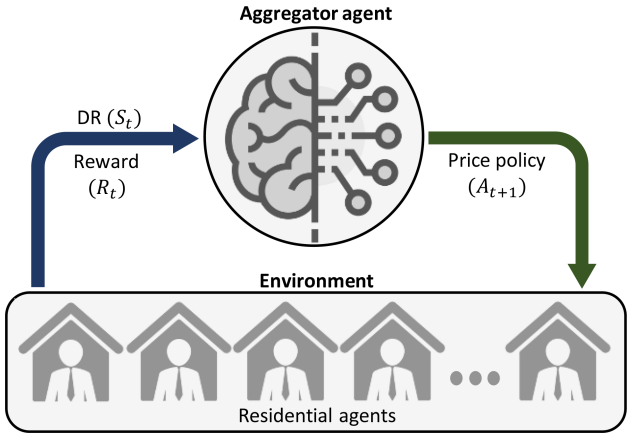


FIGURE 3. The RL environment developed using Gym toolkit.

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.

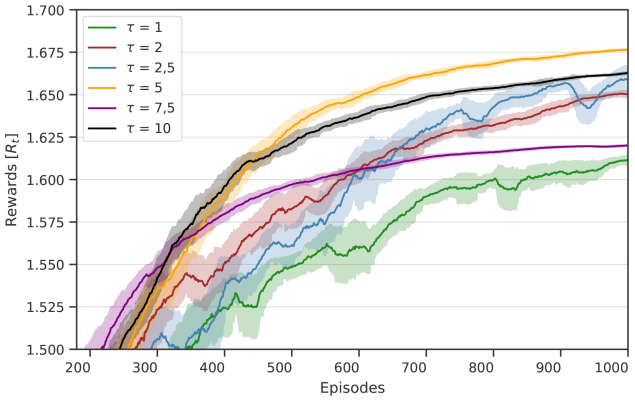


FIGURE 4. Rewards achieved by the aggregator agent within 1000 episodes of the offline training phase for different amounts of τ.

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

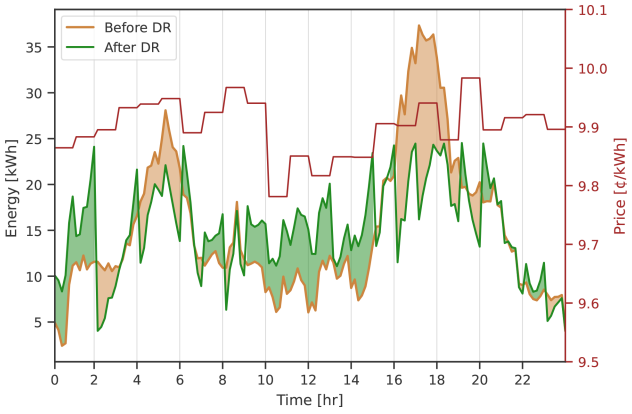


FIGURE 5. Aggregated energy consumption under the ToU tariff resulted from the offline learning process.

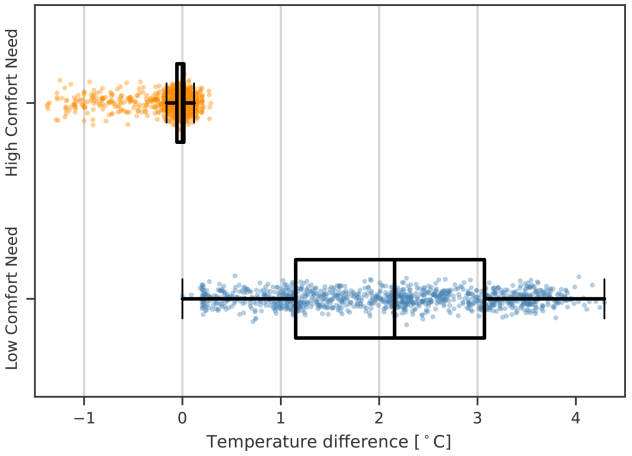


FIGURE 6. Indoor temperature deviation from set-point (temperature difference) under the DB-ToU tariff according to thermal comfort preferences of the residential agents.

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).

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

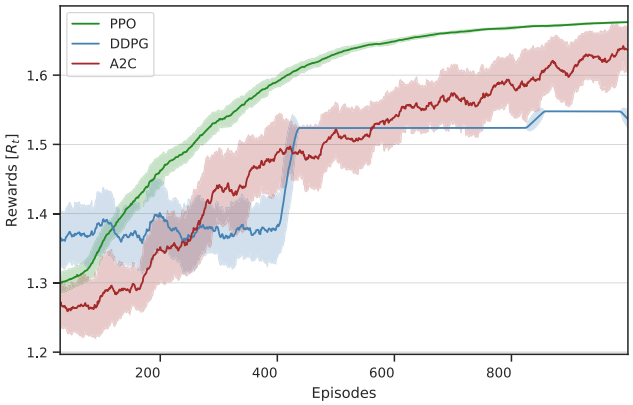
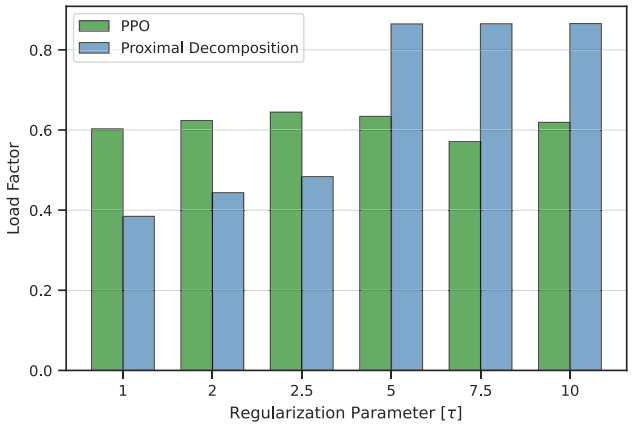
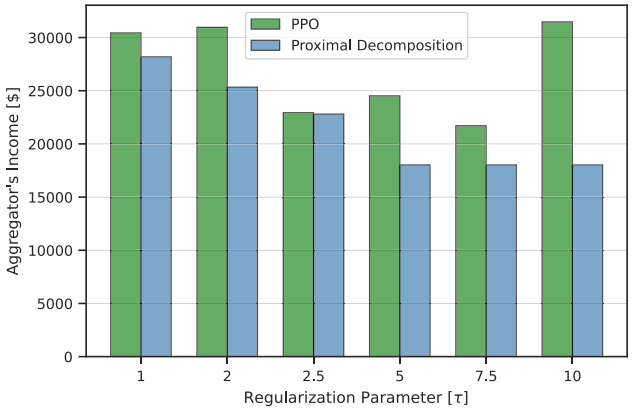


FIGURE 7. Performance comparison between different RL algorithms.



(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.

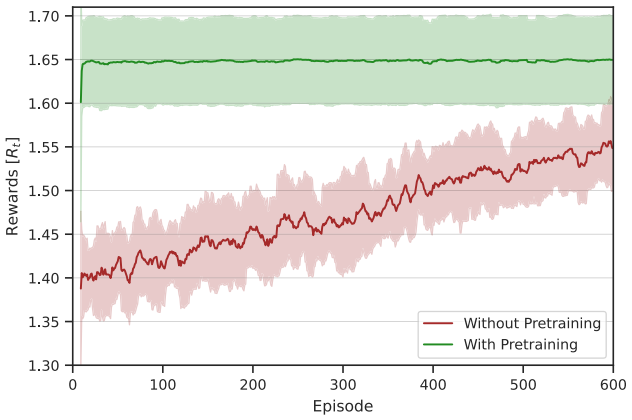


FIGURE 9. The proposed PPO performance with and without utilizing the pre-training process.

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.