# MARL-iDR

## Multi-Agent Reinforcement Learning for Incentive-Based Residential Demand Response

van Tilburg, Jasper ; Siebert, Luciano C.; Cremer, Jochen L.

**DOI**
10.1109/PowerTech55446.2023.10202941

**Publication date**
2023

**Document Version**
Final published version

**Published in**
Proceedings of the 2023 IEEE Belgrade PowerTech

*Accepted to IEEE Belgrade PowerTech, 2023*

# MARL-iDR: Multi-Agent Reinforcement Learning for Incentive-based Residential Demand Response

Jasper van Tilburg, Luciano C. Siebert, and Jochen L. Cremer
*Faculty Electrical Engineering, Mathematics & Computer Science*
*Delft University of Technology*
Delft, The Netherlands
{L.CavalcanteSiebert, J.L.Cremer}@tudelft.nl

***Abstract*—This paper presents a decentralized Multi-Agent Reinforcement Learning (MARL) approach to an incentive-based Demand Response (DR) program, which aims to maintain the capacity limits of the electricity grid and prevent grid congestion by financially incentivizing residential consumers to reduce their energy consumption. The proposed approach addresses the key challenge of coordinating heterogeneous preferences and requirements from multiple participants while preserving their privacy and minimizing financial costs for the aggregator. The participant agents use a novel Disjunctively Constrained Knapsack Problem optimization to curtail or shift the requested household appliances based on the selected demand reduction. Through case studies with electricity data from $25$ households, the proposed approach effectively reduced energy consumption's Peak-to-Average ratio (PAR) by $14.48\%$ compared to the original PAR while fully preserving participant privacy. This approach has the potential to significantly improve the efficiency and reliability of the electricity grid, making it an important contribution to the management of renewable energy resources and the growing electricity demand.**

***Index Terms*—Reinforcement Learning, Incentive-based Demand Response, Multi-Agent systems**

## I. Introduction

Demand Response (DR) initiatives are promising to satisfy the increasing need for flexibility to prevent grid congestion due to growing demands and the intermittent nature of renewable energy resources [1]. DR programs can be either price-based, where the variation in the price policy influences the demand, or incentive-based, where companies offer electricity consumers financial incentives to reduce or shift their energy consumption. Incentive-based DR (IBDR) programs are considered reward-wise programs, whereas price-based programs are considered punishment-wise programs. The voluntary nature of reward-wise programs makes people more positive and responsive in the long term. In contrast, the obligatory nature of the punishment-wise program makes people nervous, and responses are more transient [2]. IBDR programs already contribute to flexible demands in the industrial sector. Still, much less in the residential sector [3], which is a missed opportunity as residential consumers represent a significant share of electricity demand, e.g., almost half of the total energy consumption in the U.S. [4]. Moreover, residential loads can provide a more reliable and continuous response than large industrial loads [5]. However, the participation of residential consumers in IBDR programs is challenging to realize since residential participants (1) typically do not meet minimum levels of active load required to participate in the programs, (2) may not be able to respond quickly to DR events and (3) have higher privacy requirements than industrial consumers. First, existing IBDR programs are more suitable for residential participants when their loads are aggregated as a single participant in the IBDR program. Aggregators are key stakeholders in the electricity market, acting as an intermediary between the DSO and the consumer and creating the opportunity for residential consumers to participate in IBDR programs [6]. Second, requests for load reductions in IBDR program may come up unannounced and require nearly real-time response, which residential participants may not be able to manage. One approach to solve this issue is to automate the response locally at the consumer via a Home Energy Management System (HEMS). Third, to preserve the privacy of residential participants, the aggregator does not have access to detailed information about the residents' preferences. Model-based approaches to automate IBDR programs are centralized and require exhaustive information about individual participants, which may not be available or may cause privacy issues [7][8]. In addition, these centralized approaches rely on conventional optimization methods like linear programming [9][10] or dynamic programming [11], which make real-time computation infeasible for a large number of participants in the program.

Reinforcement Learning (RL) is a promising approach for decision-making in IBDR programs since it does not require any information about the organization of the program or other participants (model-free)(see Appendix A). Second, it can control multiple agents, which allows for scaling up the number of participants. Third, once trained it can decide nearly instantly, facilitating future real-time control applications.

This paper aims to answer the following research question: How can RL induce flexibility in residential demands to prevent grid congestion while preserving privacy and considering the heterogeneous preferences of residential consumers? This paper proposes a novel decentralized Multi-Agent Reinforcement Learning approach for Incentive-based DR (MARL-iDR) to answer the research question. The Markov Decision Process (MDP) is the guiding assumption to model sequential decision-making in IBDR programs. The proposed approach considers simultaneously a single Aggregator Agent (AA) and multiple

2023 IEEE Belgrade PowerTech | 978-1-6654-8778-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/POWERTECH55446.2023.10202941

participant agents aiming to maximize their rewards. The aggregator learns to deploy a suitable incentive based on one-step-ahead predictions of participant electricity demands, the target load reduction set by the DSO and participants' response to the incentive. The Participant Agent (PA) learns to respond to incentives by limiting consumption, which is achieved by shifting or curtailing household appliances, e.g. electric vehicles, dishwashers, and air conditioning, while preserving user satisfaction. The optimal power assignment is achieved through the proposed internal execution of a Disjunctively Constrained Knapsack Optimization (DCKP). This approach supports moving from inflexible centralized grid operation towards decentralized real-time automation while maintaining capacity limits and preserving consumer privacy and comfort with minimal information exchange.

The main contributions of this work are:

- An environment model that formulates an IBDR program, including an aggregator and multiple residential participants as an MDP. The environment model internally solves the DCKP to minimize participant dissatisfaction, taking the participant demand as input and schedules household appliances as output.
- MARL-iDR, a model-free MARL method for IBDR using deep Q-networks which makes real-time decisions for the aggregator and its residential participants, while preserving participants' privacy and accounting for heterogeneous preferences.

The rest of the paper is organized as follows. Section II discusses related work. The environment model is formulated as an MDP in Sec. III. In Sec. IV, the MARL-iDR algorithm is described. In Sec. V, the results of a case study are presented to test the effectiveness of the approach. Finally, Section VI concludes the paper.

## II. Related Work

Currently, much research is devoted to applying RL to DR. Some of those works focus on the industrial sector. [12] presents an approach to controlling a complex system of industrial production resources, battery storage, electricity self-supply, and short-term market trading using multi-agent RL. [13] present a deep RL-based industrial DR scheme for optimizing industrial energy management. To ensure practical application, they designed an MDP framework for industrial DR and used an actor-critic RL algorithm to determine the most efficient manufacturing schedule.

RL for DR in the residential sector has been proposed in numerous works. Many of these works focus on home energy management in a single household. [14] presents an RL-based approach to DR for a single residential or small commercial building. They apply Q-learning with eligibility traces to reduce average energy costs by shifting the time of operation of energy consuming devices either by delaying their operation or by anticipating their future use and operating them at an optimal earlier time. The algorithm balances consumer dissatisfaction with energy costs and learns consumer choices and preferences without prior knowledge about the model. [15] is a playful approach to residential DR using deep RL for scheduling loads in a single household. The authors propose an environment adapted from the Atari game Tetris where flexible blocks represent device loads. A DQN consisting of a convolutional network learns to schedule the load blocks.

Others focus on DR on the scale of the wholesale electricity market. [16] propose a voluntary incentive-based DR program targeting retail consumers with smart meters paying a flat electricity price. Load-serving entities provide consumers coupon incentives in anticipation of intermittent generation ramping and price spikes. Retail consumers' inherent flexibility is utilized while their base consumption is not exposed to wholesale real-time price fluctuations. [8] propose an incentive-based DR model considering a hierarchical electricity market including grid operators, service providers or aggregators and small-load consumers. The proposed trading framework enables system-level dispatch of DR resources by leveraging incentives between interactors. A Stackelberg game is proposed to capture the interactions between interactors.

However, the previous approaches rely on model-based algorithms instead of model-free RL. The following works propose decentralized MARL methods for load scheduling of appliances in a collection of households. [17] propose a model-free framework for scheduling the consumption profile of appliances in multiple households modelled as a non-cooperative stochastic game and apply RL to search for the Nash equilibrium. The authors emphasize the proposed method can preserve household privacy. [18] apply a cooperative RL approach to schedule controllable appliances of multiple households to minimise utility costs. The method performs explicit collaboration to satisfy global grid constraints. Both approaches emphasize the ability to scale with the number of participating households and to operate in real time.

These approaches, however, are price-based. [19] proposes a real-time RL algorithm for incentive-based DR programs that supports service providers (aggregators) to purchase energy flexibility as a resource from its subscribed residential participants to balance energy fluctuations and enhance grid reliability. A single-agent RL is adopted to compute the close-to-optimal incentive rates for heterogeneous participants. The participant's profit and dissatisfaction are balanced with the service provider's objective. [20] propose a similar method that includes PV generation and [21] propose a similar method including historical incentives.

Research on applying RL for incentive-based residential DR is scarce, and the works that address this overlap either focus on home energy management or profits from the aggregator perspective. To address this research gap, this thesis proposes a multi-agent RL algorithm for DR that is incentive-based, residential and considers the interests of both the aggregator and multiple end consumers. A comparative overview is given in Table I.

## III. Proposed Environment Model

The architecture for the proposed DR program is shown in Fig. 1, and its components are described in detail throughout

| References | Residential | Incentive-based | RL for the aggregator | RL for the consumer |
|---|---|---|---|---|
| [12], [13] | | x | | |
| [14], [15] | x | | | x |
| [16], [8] | x | x | | |
| [17], [18] | x | | | x |
| [19], [20], [21] | x | x | x | |
| MARL-IDR | x | x | x | x |

TABLE I: Overview of related work and the aspects they consider.

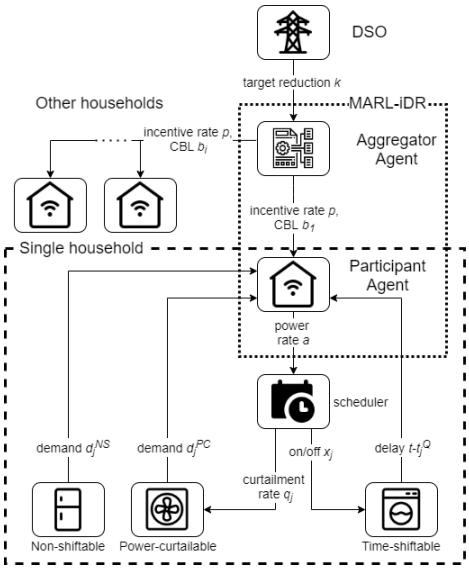



Fig. 1: Architecture of the environment

this section. The overall model considers multiple agents: one for the aggregator and one for each participant. The AA distributes incentives to the different PAs where each represents a residential household.

### A. Assumptions for DR Program and Environment Model

An assumption is that the DSO and the aggregator arrange a contractual agreement where the aggregator provides a continuous aggregated reduction in power consumption below a specified target in exchange for an agreed payment like in [22]. An additional assumption is that the participants only respond to the incentives offered in the IBDR program and not to fluctuations in the electricity price, as would be the case for price-based DR.

In IBDR programs, demand reductions are measured against a reference demand called the Customer Baseline Load (CBL). The exact demand of participants in future time steps is unknown. Hence, aggregators have to estimate the demand of their participants. Since residential participants show regular patterns in energy consumption throughout the day, the most prominent approach to demand estimation in the residential sector is based on historical consumption data. In the proposed approach, the current day is matched to ten previous similar days, and the average consumption is taken considering changes in weather conditions as the CBL. Details for calculating the CBL are found in [23].

The proposed environment model assumes an MDP, a framework for sequential decision-making, where a decision in a one-time step influences the next. The MDP is characterized by the Markov property, i.e. the state transitions depend solely on the current state of the environment and the current action taken. MDPs are described as a tuple $\langle S, A, P, R \rangle$ representing the state space, action space, transition probabilities and the reward function, respectively. One episode of an MDP consists of a finite sequence $T$ of discrete time steps $t$. The environment model assumes constant power consumption and incentive rates for a single time step. Based on these assumptions, RL agents observe the state of the environment in each time step, decide upon an action, and in return, receive a reward and transit to the next step and corresponding state. The agent considers immediate and future rewards multiplied by a discount factor $\gamma$. Therefore, the agent's objective is to maximize the cumulative discounted return [24]. However, when the reward functions are equal for multiple agents, cooperation emerges [25]. This feature is interesting to explore for multiple agents in IBDR programs.

### B. Participant Agent

The set of all households in the DR program is $\mathcal{H}$. Each PA $i \in \mathcal{H}$ can control a set of appliances $\mathcal{D}_i$. In practice, the PA could be integrated into a HEMS connected to smart meters and smart plugs to access the consumption measurements of household appliances. The objective of the PA is to approach the optimal balance between maximizing financial earnings and minimizing user dissatisfaction caused by curtailing or delaying appliances. The environment model for each household is defined by its state, the actions, rewards and the scheduler.

*State:* The appliances $\mathcal{D}_i$ are divided into three subsets: time-shiftable appliances $TS$, power-curtailable appliances $PC$ and non-shiftable appliances $NS$ such that $\mathcal{D}_i = TS_i \cup PC_i \cup NS_i$. The residents may submit an initial request to turn on appliance $j \in \mathcal{D}_i$ at time step $t_{i,j}^I$.

- Time-shiftable appliance $TS$ are either on with constant power consumption or off. Time-shiftable appliances can be interruptible (e.g. EVs, where the charging can continue later) or non-interruptible, (e.g. washing machines and dishwashers that need to complete their washing programs without interruption). Their state is determined by the difference between the current time step $t$ and the initial request time steps $t_j^I$, i.e. the current delay $t - t_j^I$.
- Power-curtailable appliances $PC$ allow to lower power consumption but do not allow a delay of usage (e.g. changing the setpoint of an AC, dimming lighting systems). Their state is the variable power demand in kilowatts $d_j^{PC}$.
- Non-shiftable appliances $NS_i$ must run at all times without delay or curtailment. These appliances share a single state defined as the total power demand in kilowatts $d_j^{NS}$

The state $s_{t,i}$ of a household $i$ at the time step $t$ combines the information of all appliances $j \in \mathcal{D}_i$. In addition, the PA observes the incentive rate $p_t$ passed from the AA and its own

projected CBL $b_{t,i}$. Hence, the observation of PA $i$ in time step $t$ is $o_{t,i}^{PA} = \{s_{t,i}, b_{t,i}, p_t\}$. [1]

*Action:* This paper proposes an action space of discrete power rates combined with an appliance scheduling optimization to ensure scalability in the number of appliances. The problem is that when appliances are controlled directly by RL the action space for time-shiftable appliances increases exponentially with the number of appliances, i.e. the binary combination of time-shiftable appliances (either on or off) is $\mathcal{O}(2^{|TS_i|})$. This problem of scalability is even more pressing for power-curtailable appliances where discretization in a set of $m$ levels of power consumption results in a combination growing with $\mathcal{O}(m^{|PC_i|})$.

To address this issue, a fixed action space is proposed that consists of discrete power rates $a \in A^{PA}$, which is a fraction of the total demand. Subsequently, the scheduler described in Section III-B matches the appliances to the limit $l = a \cdot d$ where $d$ is the total appliances' demand. The resulting total power consumption may be lower than the limit $e \leq l$.

*Reward:* The reward for the PA consists of two components (1) the financial reward for receiving incentives (2) the dissatisfaction cost for preserving the satisfaction of the residents. First, as the AA offers the PA incentive rate $p$ to reduce demand, the PA receives a financial reward when the total consumption $e$ is smaller than the CBL $b$. The financial reward $u$ paid from AA to PA is

$$u = p \cdot \max(0,\, b - e), \tag{1}$$

As the DR program is incentive-based (reward-wise) not price-based (punishment-wise), participants are not punished for consuming more than CBL $b$, i.e. they can only earn money, not lose anything.

Second, curtailing or shifting requested appliances causes dissatisfaction to the residents. In the case of time-shiftable appliance $j \in TS$, dissatisfaction cost $c_j^{TS}$ is a convex function of the delay

$$c_j^{TS} = \beta_j \Big(t + 1 - t_j^I\Big)^2 \quad \forall j \in TS, \tag{2}$$

Shifting appliance $j$ to time step $t+1$ instead of turning it on in time step $t$ means a delay of $t+1-t_j^I$. This function assumes the residents get increasingly dissatisfied when waiting longer for the appliance to run [26]. In the case of power-curtailable appliance $j \in PC$, the dissatisfaction cost is a convex function of the power curtailment.

$$c_j^{PC} = \beta_j \Big(\frac{1}{m} \cdot q_j \cdot d_j\Big)^2 \quad \forall j \in PC, \tag{3}$$

where $q_j \in \{0, 1, \ldots, m\}$ is a categorical variable corresponding to the power curtailment level. This function assumes residents get increasingly dissatisfied with increased curtailment [27]. $\beta_j$ is an appliance-specific dissatisfaction coefficient describing the tolerance of the residents for delay or power curtailment. In practice, this coefficient is a parameter that the residents can update in the HEMS according to their preferences.

The total reward function combines financial reward $u$ and dissatisfaction cost $c$ as follows

$$r^{PA} = u - \sum_{j \in \mathcal{D}} c_j \tag{4}$$

*Scheduler:* As part of the PA for household $i$, the scheduler determines the optimal assignment of power to the appliances based on the overall demand limit $l$. The scheduler is a combinatorial optimization formulated as DCKP [28]:

$$\text{minimize} \quad \sum_{j \in PC} c_j^{PC} + \sum_{j \in TS} (1 - x_j) \cdot c_j^{TS} \tag{5}$$

$$\text{subject to} \quad \sum_{j \in PC} \frac{1}{m} \cdot q_j \cdot d_j + \sum_{j \in TS} x_j \cdot d_j \leq l \tag{6}$$

$$x \in \{0, 1\},\, q \in \{0, 1, \ldots, m\} \tag{7}$$

where $x_j$ is a binary variable for each time-shiftable appliance $j \in TS$ corresponding to switching the appliance off ($x_j = 0$) or on ($x_j = 1$). The optimization minimizes the total dissatisfaction from all appliances. The dissatisfaction costs from time-shiftable appliances $c_j^{TS}$ and power-curtailable appliances $c_j^{PC}$ are parameters computed with Eq. (2) and Eq. (3). After solving the DCKP, the overall power demand is

$$e = \sum_{j \in PC} \frac{1}{m} \cdot q_j \cdot d_j + \sum_{j \in TS} x_j \cdot d_j \tag{8}$$

### C. Aggregator Agent

*State:* The state space of the AA is $o_t^{AA} = \{d_t, k\}$, where $d_t$ is the aggregated demand of all households $i \in \mathcal{H}$ and $k$ is the target reduction set by the DSO.

*Action:* In each time step the AA selects an incentive rate $p_t$ to realize power reduction by the PAs. The AA selects $p_t$ out of an action space of discrete incentives $A^{AA}$ in cents per kilowatt of demand reduction.

*Reward:* The reward of the AA is

$$r_t^{AA} = -\Big(\rho \cdot e_t^+ + (1 - \rho) \cdot \sum_{i \in \mathcal{H}} u_{t,i}\Big), \tag{9}$$

where the first term defines a penalty for exceeding target $k$ as $e_t^+ = max(0, e_t - k)$. The second term is the total incentive paid to the PAs as defined in Eq. (1). The trade-off between the two terms is determined by weighting factor $\rho$. Note that the AA is not rewarded for aggregated consumption below the target as it aims to reduce consumption to contribute to the DSO's capacity constraints, but not further reduce energy consumption.

## IV. Proposed MARL-iDR Algorithm

The proposed MARL-iDR is a multi-agent algorithm where the AA and PA have indirectly opposing reward functions, i.e. actions in favour of the AA may have a negative influence on the reward of the PA and vice versa. All MARL-iDR agents are trained simultaneously, hence, the agents deal with a moving target where the optimal policy changes as opposing

[1] For the remainder of this subsection, the subscripts $t$ and $i$ are dropped as all equations apply to time step $t$ and household $i$.

agents change their policies. Simultaneous learning leads to non-stationary problems which invalidate most of the single-agent RL theoretical guarantees, e.g. the guarantee of convergence [25]. Despite these limitations, simultaneous learning has found numerous applications because of its simplicity [29][30].

MARL-iDR effectively trades off exploration with exploitation (a fundamental concept in RL). MARL-iDR uses the action-selection strategy of $\epsilon$-greedy with decay, i.e. with probability $\epsilon$ the agent selects a random action where $\epsilon$ decreases over time with decay rate $\delta$ [31]. With this strategy, MARL-iDR benefits from extensive exploration early in training and refinement of the policy in later stages.

MARL-iDR uses Deep Q-Networks (DQNs) to account for huge and continuous state spaces that are infeasible to Q-learning [24]. DQN is a state-of-the-art Deep RL approach that estimates the Q-value of state-action pairs by means of a neural network $\theta$. For more stable training with DQNs, two features are used: 1) A separate target network $\theta^T$ for setting the target values to avoid non-stationary targets, while the original network is used for predicting Q-values. 2) Experience replay to avoid the agent from forgetting previous experiences. All experiences are saved in a replay buffer $B$. Instead of training the network only on the most recent experience, the network is trained on randomly sampled batches of experience from $B$.

The training procedure for MARL-iDR is Algorithm 1. At the start of the procedure a policy network and target network are initialized for each individual agent. Then, all agents train the networks for a number of episodes where each episode corresponds to a single day which has a sequence of $T$ time steps. In each time step $t$, first the AA selects an action. Next, each individual PA $i$ selects an action and immediately receive their reward. Finally, after all PAs decided their response to the AA the reward for the AA can be calculated. The training procedure takes a significant amount of time to learn policies for each agent, however, once trained, the RL agent can be deployed in real-time using policy $\pi$:

$$\pi(s) = \underset{a}{argmax}\, Q(s, a \,|\, \theta) \tag{10}$$

**Algorithm 1** MARL-iDR training procedure

```
Initialize θ^AA, θ^{T,AA}, B^AA
Initialize θ_i^PA, θ_i^{T,PA}, B_i^PA    ∀i ∈ H
Initialize ε_0 ← 1.0
Initialize δ,   0 < δ < 1
for all episodes do
    Initialize target reduction k
    Initialize rewards r_0^AA, r_0^PA ← 0
    ε_t = ε_{t−1} · δ
    for all time steps t do
        Predict demand d_t and set CBLs b_t
        o_t^AA ← ⟨d_t, k⟩
        Add ⟨o_{t−1}^AA, p_{t−1}, r_{t−1}^AA, o_t^AA⟩ to B
        Train θ^AA given B
        Select p_t using ε-greedy
        for all PAs i ∈ H do
            Observe state of appliances s_{t,i} and compute c_{t,i}
            Observe CBL b_{t,i}, incentive rate p_t
            o_{t,i}^PA ← ⟨s_{t,i}, b_{t,i}, p_t⟩
            Add ⟨o_{t−1,i}^PA, a_{t−1,i}, r_{t−1,i}^PA, o_t^PA⟩ to B
            Train θ_i^PA given B
            Select a_{t,i} using ε-greedy
            Obtain x_{t,i} and q_{t,i} by solving DCKP
                with input: d_{t,i}, c_{t,i}, a_{t,i}
            Update s_{t+1,i} according to x_{t,i} and q_{t,i}
            Calculate PA reward r_{t,i}^PA
        end for
        Calculate AA reward r_t^AA
    end for
end for
```

| Appliance | Type | Demand (kW) | Dissatisfaction coeff. mean | std |
|---|---|---|---|---|
| Dryer | $TS$, $NI$ | 2.0 | 0.2 | 0.2 |
| Washing machine (WM) | $TS$, $NI$ | 1.0 | 0.1 | 0.1 |
| Dishwasher (DW) | $TS$, $NI$ | 2.0 | 0.06 | 0.05 |
| EV | $TS$, $I$ | 4.0 | 0.04 | 0.05 |
| AC | $PC$ | 0 - 4.0 | 3.0 | 1.0 |
| Non-shiftable | $NS$ | 0 - 5.0 | - | - |

TABLE II: Household appliances and their parameters, e.g, non-interruptible ($NI$) and interruptible ($I$).

## V. Case study

The case study tests the effectiveness of the environment model and the MARL-iDR algorithm taking four aspects into consideration: 1) The policies learned by the agents 2) Information exchange and if the privacy of the participants is preserved 3) Computational efficiency and finally 4) Economics for the aggregator considering a varying weighting factor $\rho$.

### A. Simulation data and test setup

This case study uses appliance requests and consumption data from 25 real-world households from the PecanStreet dataset [32]. The dissatisfaction coefficients $\beta_j$ for the appliances are sampled from a normal distribution to introduce heterogeneity to the households. The type, demand and dissatisfaction coefficients of the appliances selected for the simulations are in Table II.

The MARL-iDR algorithm is trained for $5000$ episodes, discretized in $T = 96$ time steps of $15\,\mathrm{min}$ and randomly sampled from the training period April 1, 2018, to October 31, 2018. The action space for the PAs and the AAs are defined as $A^{PA} = \{0.0, 0.1, ..., 1.0\}$ and $A^{AA} = \{0, 1, ..., 10\}$ respectively. The scheduler selected from $m = 10$ curtailment levels. The AA and PAs have an individual DQN with a learning rate $\eta = 0.001$. The discount rate $\gamma = 0.9$ and $\epsilon$-decay rate $\delta = 0.999$. Weighting factor $\rho$ is $0.5$. Finally, the target reduction $k$ is defined at $80\%$ of the peak demand. The algorithm is available online in a GitHub repository [33]. The

| | No DR | MARL-IDR | Myopic baseline |
|---|---|---|---|
| Peak load (kW) | 86.25 | 74.39 | 69.23 |
| Mean load (kW) | 47.80 | 45.37 | 46.23 |
| PAR | 1.80 | 1.64 | 1.50 |
| Surplus consumption (kWh) | 35.79 | 3.93 | 0.49 |
| Total incentive (¢) | 0.0 | 2122 | 1917 |
| Average dissatisfaction cost | 0.0 | 17.36 | 12.26 |
| Average incentive income (¢) | 0.0 | 84.88 | 76.68 |

TABLE III: Results averaged per day in July.

algorithm is validated on each day in July. The simulations were conducted on a 2.20 GHz, Intel 6-core i7-8750 CPU with 16 GB RAM, running Windows 10.

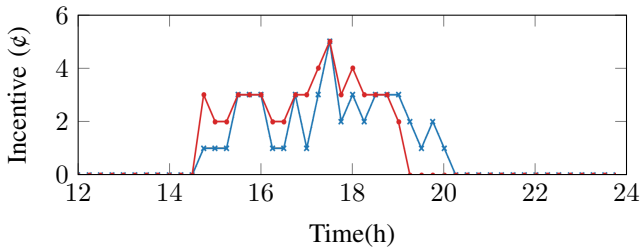
A baseline was used to compare the performance of the proposed MARL-iDR. The baseline considered the optimal myopic action per time step (i.e. not considering future rewards). In other words, PAs selected the best action such that $a_i^* = argmax\,\{r_i^{PA}|p\}$, and the AA selected the optimal incentive defined by $p^* = argmax\,\{r^{AA}|a_i^*, \forall i \in H\}$. This myopic baseline requires full model knowledge and can only consider immediate rewards (short-sighted).

### B. Load reductions and incentive rates

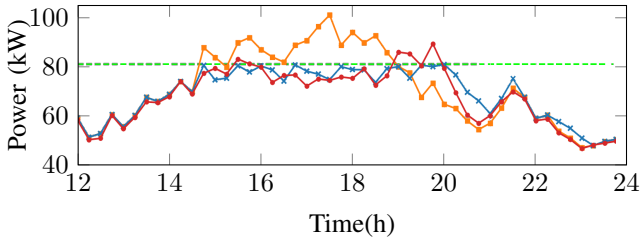
MARL-iDR reduces loads during peak hours. The results are in Table III. The peak load and peak-to-average ratio (PAR) are significantly lower for MARL-iDR compared to the original load, i.e the case without DR. However, the myopic baseline reduces, even more, the peak load and PAR, slightly exceeding the target reduction with a total of $0.49\,\mathrm{kWh}$, whereas this "surplus consumption" for MARL-iDR is $3.93\,\mathrm{kWh}$. MARL-iDR sometimes results in a second peak that exceeds the target reduction $k$. This behaviour of shifting the load to a second peak is known as the rebound effect [34]. Fig. 2 illustrates this behaviour, showing the impact of MARL-iDR on the load curve on July 1st. The aggregated load (Fig. 2b) is unchanged before 14:30. During hours where the original load exceeds the target reduction, both MARL-iDR and the myopic baseline maintained the total load mostly below the target, by offering varying incentive rates. However, around 19:00, a second peak arises when using MARL-iDR. This second peak of $90.7\,\mathrm{kW}$ is lower than the first, original peak $102.7\,\mathrm{kW}$ but higher than the target. MARL-iDR does not offer incentives after the original peak. Hence the loads increase above the target (rebound effect). The myopic baseline does not suffer from the rebound effect and reduces below the target. A similar pattern can be observed in individual households, see Fig. 2c for the load curve of a selected household. Similar to the aggregated case, consumption is reduced significantly during peak hours but spikes right after 19:00.

### C. Dissatisfaction costs and appliance scheduling

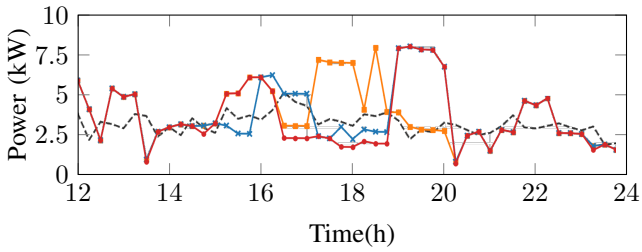
The appliance schedule of the household from Fig. 2c is in Fig. 3. Fig. 3a shows each appliance's originally requested time step and the scheduled time step. The washing machine




(a) Incentive rates selected by the AA.




(b) Aggregated load curve with target reduction (▪▪).




(c) Load curve of an individual household with CBL (▪▪).

Fig. 2: Load reductions and incentive rates for MARL-iDR (▬) compared to the myopic baseline (▬) and without DR (▬).

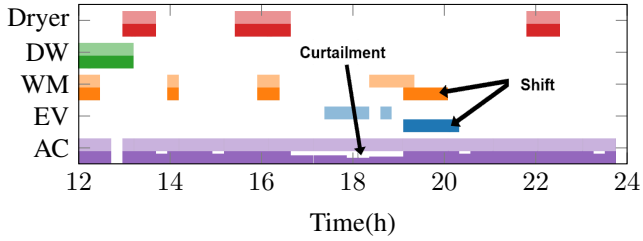
and the EV delay for as long as incentives are offered. As soon as the incentive rate drops to 0, shortly after 19:00, the agent schedules the washing machine and the EV. The incentive also influences the AC. Between 16:30 and 19:00, the AC consumption is nearly halved. Fig. 3b shows the trade-off between incentives gained and dissatisfaction caused by rescheduling appliances. The dissatisfaction from postponing EV charging and the washing machine increases until 19.15.

### D. Preserving privacy

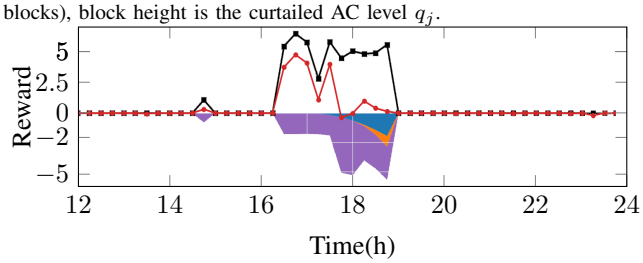
The proposed MARL-iDR preserves privacy which is legally required. MARL-iDR outperforms any centralized scheduler at the AA (e.g., myopic baseline) as they require knowing all resident details, such as exact information about the reward function, to calculate the best responses. In practice, the aggregator must know participants' preferences and the state of their appliances to predict their response. With MARL-iDR the aggregator receives no information regarding the participant; only a single value-information is exchanged, the incentive rate. In addition, no information is exchanged between participants, ensuring participant privacy.

### E. Economic analysis

MARL-iDR trains local agents to balance the economics between the aggregator and all participants. The key metric




(a) Schedule with requests (light blocks) and assignments (dark blocks), block height is the curtailed AC level $q_j$.




(b) Trading-off the two reward components of the PA, the financial reward $u$ minus the accumulated dissatisfaction $c$.

Fig. 3: Appliance scheduling of one household in (a) and in (b) is the corresponding financial reward $u$ (—■—), total reward $r^{PA}$ (—●—) and all components to the dissatisfaction $c_{EV}$ (■), $c_{WM}$ (■), $c_{DW}$ (■), $c_{Dryer}$ (■), $c_{AC}$ (■).

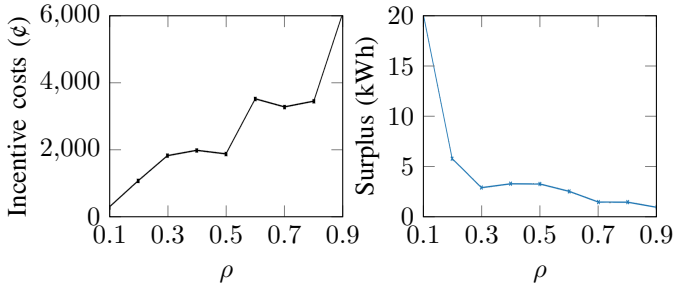
in this balance is the incentive rate and the AA-designed parameters of the DR scheme.

The incentive rates selected by the AA are shown in Fig. 2a. The shape of this rate curve matches the original aggregated demand curve in Fig. 2b. MARL-iDR stops offering incentives after 19.00, and the myopic baseline stops after 20.00. The myopic baseline offers less or equal incentives optimizing the target reduction between 14.45 and 18.15, which results in lower financial rewards for the PAs.

The design of the program is important for its success toward a fair balance in financial costs and gains for all parties. The design (and the balance) is controlled at the AA by selecting the weighting factor $\rho$ in Eq. (9). The selection of $\rho$ ultimately determines households' willingness to participate in the DR program. A study on this parameter $\rho$ is in Fig. 4. The larger $\rho$, the larger the punishment on surplus consumption and the smaller the incentive cost. For small values $\rho$, a significant consumption exceeds the target while only a little incentive is paid to participants. On the other hand, when $\rho$ is large, the aggregator tries to push the surplus consumption down to 0 by offering increasing incentives.

### *F. Computational efficiency*

This study analyzes the computational times for MARL-iDR training and real-time deployment. The computation time during real-time deployment is an important criterion for the future needs of residential DR programs. The computation times of scheduling the appliances of a single household in a one-time step are compared with the baseline. A myopic baseline is a centralized approach to computing all possibilities before the PA can decide the optimal scheduling. The used implementation took, on average, 1.86 s, and for




(a) (b)

Fig. 4: The aggregator balances financial cost (a) and surplus consumption (b) with the parameter $\rho$.

a large number of households, centralized optimization-based approaches are highly unsuitable, as research shows. However, as MARL-iDR is decentralized and the PAs can schedule appliances independently, the actual schedule can be computed in 2 ms. As the scheduling of appliances should be done in near real-time, MARL-iDR is very suitable as a real-time decision-making algorithm for real-time DR programs. The key advantage is that almost unlimited many PAs can be considered simultaneously. The time of training MARL-iDR with one AA and 25 PAs for 5000 episodes is $\sim$ 12 h, which only has to be done once before deployment.

The proposed decentralized MARL-iDR approach is very promising for future real-time DR programs as it scales to very large numbers of residents, making DR decisions in milliseconds while preserving privacy and balancing financial gains among participants in a fair way. However, MARL-iDR has limitations. As the AA makes decisions based on the current state of the environment and can not know if the current time step is before or after the peak as of the nature of MDPs, incentives were not placed to reduce the second peak. Hence, the myopic baseline outperformed MARL-iDR in reducing load. One way of solving this limitation could be to include the accumulated load reduction in the observation which requires further investigation. Another limitation is that the scheduler only considers requests for appliances at time step $t$, hence non-interruptible appliances may impede load reduction in future time steps when incentives may be higher. Operation times of time-shiftable appliances must be considered in the future to improve the potency of appliance scheduling. Finally, MARL-iDR is trained and validated in a period with relative high outside temperatures and large AC consumption (April to October). In the future, different characteristics should be considered to analyze the generalizability of the proposed method.

## VI. Conclusion

In conclusion, this paper proposed a decentralized Multi-Agent Reinforcement Learning (MARL) approach to an incentive-based Demand Response (DR) program that addresses the key challenge of coordinating heterogeneous preferences and requirements from multiple participants while

preserving their privacy and minimizing financial costs for the aggregator. The proposed approach was validated through case studies with electricity data from 25 households. It was shown to effectively reduce the Peak-to-Average ratio (PAR) of energy consumption by 14.48% compared to the original PAR while fully preserving participant privacy. However, the MARL-IDR algorithm showed some rebound effects and did not always achieve the target reduction and the myopic baseline. The results of this case study demonstrate the proposed approach's potential to improve the electricity grid's efficiency and reliability. The novel Disjunctively Constrained Knapsack Problem optimization used to curtail or shift the requested household appliances based on the selected demand reduction makes this approach valuable to managing renewable energy resources and the growing electricity demand. Future work should address the rebound effect and improve the algorithm's performance.

## REFERENCES


[1] D. Li, W. Y. Chiu, and H. Sun, "Demand Side Management in Microgrid Control Systems," in *Microgrid: Advanced Control Methods and Renewable Energy System Integration*. Elsevier Inc., 1 2017, pp. 203–230.

[2] P. T. Baboli, M. Eghbal, M. P. Moghaddam, and H. Aalami, "Customer behavior based demand response model," *IEEE Power and Energy Society General Meeting*, 2012.

[3] Z. Wang, H. Li, N. Deng, K. Cheng, B. Lu, B. Zhang, and B. Wang, "How to effectively implement an incentive-based residential electricity demand response policy? Experience from large-scale trials and matching questionnaires," *Energy Policy*, vol. 141, p. 111450, 6 2020.

[4] "Annual Energy Outlook 2021." [Online]. Available: https://www.eia.gov/outlooks/aeo/

[5] A. Asadinejad, K. Tomsovic, and C. F. Chen, "Sensitivity of incentive based demand response program to residential customer elasticity," in *NAPS 2016 - 48th North American Power Symposium, Proceedings*. Institute of Electrical and Electronics Engineers Inc., 11 2016.

[6] K. Zhou and S. Yang, "Smart Energy Management," in *Comprehensive Energy Systems*. Elsevier, 1 2018, vol. 5-5, pp. 423–456.

[7] S. Maharjan, Q. Zhu, Y. Zhang, S. Gjessing, and T. Başsar, "Dependable demand response management in the smart grid: A stackelberg game approach," *IEEE Transactions on Smart Grid*, vol. 4, no. 1, pp. 120–132, 2013.

[8] M. Yu and S. H. Hong, "Incentive-based demand response considering hierarchical electricity market: A Stackelberg game approach," *Applied Energy*, vol. 203, pp. 267–279, 10 2017.

[9] A. Gholian, H. Mohsenian-Rad, and Y. Hua, "Optimal Industrial Load Control in Smart Grid," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2305–2316, 9 2016.

[10] P. R. Senevirathne, L. R. Senarathne, I. U. Muthunaike, J. V. Wijayakulasooriya, U. S. Nawaratne, and H. A. Dharmagunawardana, "Optimal Residential Load Scheduling in Dynamic Tariff Environment," *ICIIS - Proceedings*, pp. 547–552, 12 2019.

[11] Y. J. A. Zhang, C. Zhao, W. Tang, and S. H. Low, "Profit-maximizing planning and control of battery energy storage systems for primary frequency control," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 712–723, 2018.

[12] M. Roesch, C. Linder, R. Zimmermann, A. Rudolf, A. Hohmann, and G. Reinhart, "Smart grid for industry using multi-agent reinforcement learning," *Applied Sciences (Switzerland)*, vol. 10, no. 19, pp. 1–20, 10 2020.

[13] X. Huang, S. H. Hong, M. Yu, Y. DIng, and J. Jiang, "Demand Response Management for Industrial Facilities: A Deep Reinforcement Learning Approach," *IEEE Access*, vol. 7, pp. 82 194–82 205, 2019.

[14] Z. Wen, D. O'Neill, and H. R. Maei, "Optimal Demand Response Using Device Based Reinforcement Learning," *IEEE Transactions on Smart Grid*, vol. 6, no. 5, pp. 2312–2324, 1 2014. [Online]. Available: https://arxiv.org/abs/1401.1549v2

[15] A. Mathew, A. Roy, and J. Mathew, "Intelligent residential energy management system using deep reinforcement learning," *IEEE Systems Journal*, vol. 14, no. 4, pp. 5362–5372, 2020.

[16] H. Zhong, L. Xie, and Q. Xia, "Coupon incentive-based demand response: Theory and case study," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1266–1276, 2013.

[17] H.-M. Chung, S. Maharjan, Y. Zhang, and F. Eliassen, "Distributed deep reinforcement learning for intelligent load scheduling in residential smart grids," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2752–2763, 2021.

[18] C. Zhang, S. R. Kuppannagari, C. Xiong, R. Kannan, and V. K. Prasanna, "A cooperative multi-agent deep reinforcement learning framework for real-time residential load scheduling," *IoTDI 2019 - Proceedings of the 2019 Internet of Things Design and Implementation*, pp. 59–69, 4 2019.

[19] R. Lu and S. H. Hong, "Incentive-based demand response for smart grid with reinforcement learning and deep neural network," *Applied Energy*, vol. 236, pp. 937–949, 2 2019.

[20] L. Wen, K. Zhou, J. Li, and S. Wang, "Modified deep learning and reinforcement learning for an incentive-based demand response model," *Energy*, vol. 205, p. 118019, 8 2020.

[21] H. Xu, W. Kuang, J. Lu, and Q. Hu, "A Modified Incentive-based Demand Response Model using Deep Reinforcement Learning," in *Asia-Pacific Power and Energy Engineering Conference, APPEEC*, vol. 2020-September. IEEE Computer Society, 9 2020.

[22] "Base Interruptible Program (BIP)." [Online]. Available: https://www.pge.com/

[23] R. Sharifi, S. H. Fathi, and V. Vahidinasab, "Customer baseline load models for residential sector in a smart-grid environment," *Energy Reports*, vol. 2, pp. 74–81, 11 2016.

[24] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction Second edition*. The MIT Press, 2018.

[25] L. Buşoniu, R. Babuška, and B. De Schutter, "Multi-agent reinforcement learning: An overview," *Studies in Computational Intelligence*, vol. 310, pp. 183–221, 2010.

[26] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, and C. S. Lai, "A Multi-Agent Reinforcement Learning-Based Data-Driven Method for Home Energy Management," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3201–3211, 7 2020.

[27] M. Fahrioglu and F. L. Alvarado, "Using utility information to calibrate customer demand management behavior models," *IEEE Transactions on Power Systems*, vol. 16, no. 2, pp. 317–323, 5 2001.

[28] M. Ben Salem, R. Taktak, A. R. Mahjoub, and H. Ben-Abdallah, "Optimization algorithms for the disjunctively constrained knapsack problem," *Soft Computing*, vol. 22, no. 6, pp. 2025–2043, 12 2016. [Online]. Available: https://link-springer-com.tudelft.idm.oclc.org/article/10.1007/s00500-016-2465-7

[29] R. Crites and A. Barto, "Improving Elevator Performance Using Reinforcement Learning," *Advances in Neural Information Processing Systems*, vol. 8, 1995.

[30] M. J. Matarić, "Learning in multi-robot systems," *Lecture Notes in Computer Science*, vol. 1042, pp. 152–163, 1995.

[31] O. Caelen and G. Bontempi, "Improving the Exploration Strategy in Bandit Algorithms," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5313 LNCS, pp. 56–68, 2007.

[32] "Pecan Street Inc." [Online]. Available: https://www.pecanstreet.org/

[33] J. van Tilburg, L. C. Siebert, and J. L. Cremer, "Case study for MARL-iDR-Multi-Agent-Reinforcement-Learning-for-Incentive-based-Residential-Demand-Response." [Online]. Available: https://github.com/TU-Delft-AI-Energy-Lab/MARL-iDR-Multi-Agent-Reinforcement-Learning-for-Incentive-based-Residential-Demand-Response

[34] Z. Broka and K. Baltputnis, "Handling of the Rebound Effect in Independent Aggregator Framework," *International Conference on the European Energy Market, EEM*, 9 2020.