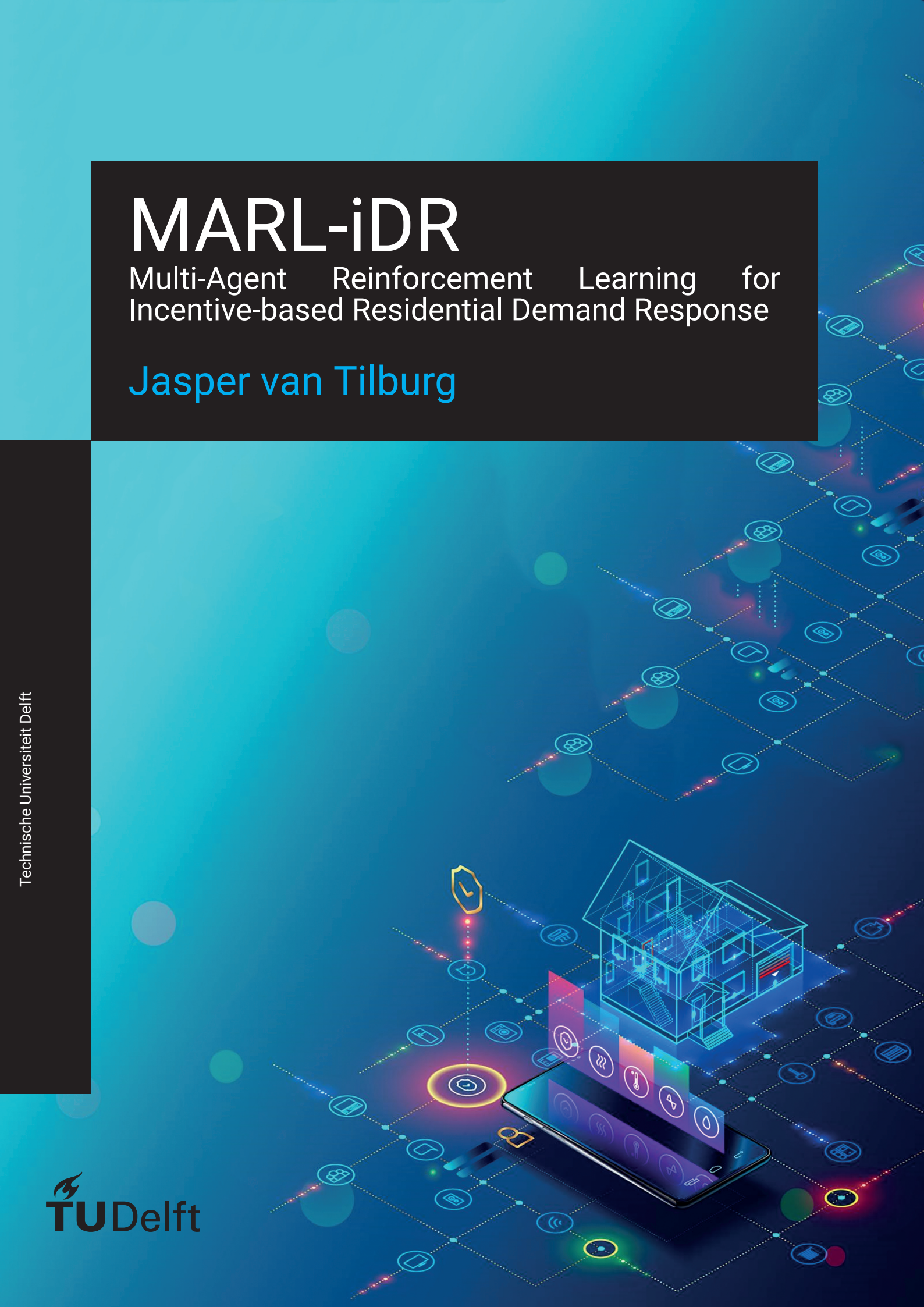
MARL-iDR

Multi-Agent Reinforcement Learning for Incentive-based Residential Demand Response

Jasper van Tilburg

Technische Universiteit Delft



TU Delft

MARL-iDR

Multi-Agent Reinforcement Learning for Incentive-based Residential Demand Response

by

Jasper van Tilburg

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on Tuesday, November 30, 2021 at 9:00 AM.

Student number:	4393554	
Project duration:	February 16, 2021 – November 30, 2021	
Thesis committee:	Dr. L. C. Siebert,	TU Delft (supervisor)
	Dr. J.L. Cremer,	TU Delft (co-supervisor)
	Prof. dr. C. M. Jonker,	TU Delft
	Dr. P. V. Barrios,	TU Delft

This thesis is confidential and cannot be made public until November 30, 2021.

TU Delft

The work in this thesis has been submitted as a scientific paper to the 2022 Power Systems Computation Conference. If the article is accepted, it will be published in a special issue of the journal Electric Power Systems Research indexed by ScienceDirect.

Abstract

Distribution System Operators (DSOs) are responsible preventing grid congestion, while accounting for growing demand and the intermittent nature of renewable energy resources. Incentive-based demand response programs promise real-time flexibility to relieve grid congestion. To include residential consumers in these programs, aggregators can financially incentivize participants to reduce their energy demand and make aggregated energy reduction available to DSOs. A key challenge for aggregators is to coordinate heterogeneous preferences from multiple participants while preserving their privacy. This thesis proposes MARL-iDR: a decentralized Multi-Agent Reinforcement Learning approach to an incentive-based demand response program. The approach respects participants' privacy and preferences and makes decisions in real-time when deployed. The aggregator and each participant are controlled by Deep Reinforcement Learning agents that learn to maximize their reward. The aggregator agent learns a policy that dispatches suitable incentives to participants based on total energy demand and a target reduction, while minimizing financial costs. The participant agent learns to respond to these incentives by reducing consumption to a fraction of the original demand. The participant agents curtail or shift requested household appliances based on the selected consumption reduction using a novel Disjunctively Constrained Knapsack Problem optimization, while minimizing residents' dissatisfaction. A case study with real-world electricity data from 25 households demonstrates the capability to induce demand-side flexibility. The approach is compared to the case without demand response and to a centralized myopic baseline approach. A 9% reduction of the Peak-to-Average ratio (PAR) was achieved compared to the original PAR (no demand response).

Preface

Imagine a young boy learning to juggle or a newborn deer calf taking its first steps. Both undergo great efforts in pursue of a rewarding goal. In real life, Reinforcement Learning is everywhere. Everywhere around us we see the tedious, yet wonderful process of learning through experience. No wonder it inspired researchers to explore a scientific approach to imitate this complex learning mechanism. Since the introduction of Reinforcement Learning in the world of computer science, studies have gained success after success on both theoretical as well as practical fronts. However, the delicate process of trial and error is at the same time the biggest power of the Reinforcement Learning paradigm and its biggest limitation. For example, it would cost thousands of cars before computers learn to drive one without trashing it. Given the proper training in a suitable simulation, Reinforcement Learning has proved extremely successful and its application has provided breakthroughs in healthcare, finance, manufacturing, robotics and, not to forget power systems.

In the 21 century, it is an enormous challenge to balance the supply and demand of electricity. To avoid overload and shortages, flexibility on the demand side has become an important resource. Demand Response is an initiative whereby households reduce their electricity consumption in exchange for financial compensation. In my thesis, I investigate the potential of automating Demand Response programs by means of Reinforcement Learning. Think of computation of the correct financial incentive on the one hand, and automated management of household appliances on the other. An approach beneficial to both grid operators and end consumers. In the future, this approach may help utility companies supply us with our daily electricity needs.

I want to express my gratitude towards Dr. Luciano Cavalcante Siebert, my supervisor, who guided me through this learning experience, and towards Dr. Jochen Cremer, whose thoughtful and critical inputs have led to inspiring discussions.

Jasper van Tilburg
Delft, November 2021

Nomenclature

Acronyms

AA	Aggregator Agent
CBL	Customer Baseline Load
DCKP	Disjunctively Constrained Knapsack Problem
DQL	Deep Q-Learning
DQN	Deep Q-Network
DR	Demand Response
DSO	Distribution System Operator
HEMS	Home Energy Management System
IBDR	Incentive-Based Demand Response
MARL	Multi-Agent Reinforcement Learning
MDP	Markov Decision Process
PA	Participant Agent
PBDR	Price-Based Demand Response
RL	Reinforcement Learning

Symbols

$\beta_{i,j}$	Dissatisfaction coefficient of appliance j of PA i
δ	Decay rate of ϵ
ϵ	Probability of taking a random action
γ	Discount factor
NS	Set of non-shiftable appliances
PC	Set of power-curtailable appliances
TS	Set of time-shiftable appliances
$b_{t,i}$	CBL of PA i in time step t
$c_{t,i,j}$	Dissatisfaction cost of appliance j of PA i in time step t
$d_{t,i}$	Total demand of PA i in time step t
e_t^+	Surplus consumption in time step t, i.e., power exceeding k
$i/\mathcal{H}$	Index/set of households/PAs
$j/\mathcal{D}$	Index/set of household appliances

k	Target reduction for the AA
m	Amount of categorical levels of power curtailment
o_t^{AA}	Observation of the AA in time step t
$o_{t,i}^{PA}$	Observation of PA i in time step t
p_t	Incentive set by the AA in time step t
q_j	Categorical level of power curtailment for appliance j
r_t^{AA}	Total reward for the AA in time step t
$r_{t,i}^{PA}$	Total reward for PA i in time step t
$s_{t,i,j}$	State of appliance j of PA i in time step t
t/T	Index/set of time steps
$t_{i,j}^{I}$	Time step in which appliance j of PA i is requested
$u_{t,i}$	Financial reward of PA i in time step t

Contents

1 Introduction

Historically, the demand side of the electricity market is considered inelastic and the generation side should adapt to fully supply the demand. Grid operators were able to match the demand by throttling the production rate of conventional power sources, like coal and oil plants [5]. However, firing up additional fossil fuel power plants is expensive and takes some time to come up to full production. Moreover, in pursue of climate goals, renewable energy sources are phasing out fossil fuels. Renewable energy sources are intermittent and can be unreliable, e.g., wind and solar power generation depends on the weather [33]. The rise of renewable energy sources, together with increasing electricity demand, poses challenges to grid operators when matching supply and demand, i.e., grid balancing. A particular challenge in grid balancing is congestion management. The Distribution System Operator (DSO) is a utility company (utility in short) responsible for distributing and managing electricity from substations to the end consumers while maintaining the capacity limits of the grid. When these limits are exceeded we call it a grid congestion. Grid congestion can be costly or even cause power outages.

Flexibility in the demand side has been recognized as an important resource to aid DSOs in congestion management [34]. Demand Response initiatives are promising to satisfy this need for flexibility. Demand Response (DR) refers to a change in consumption patterns of electricity customers in response to financial incentives in times when the network reliability is at risk [15]. DR programs are divided into price-based DR (PBDR) schemes, e.g., time-of-use pricing, critical-peak pricing and real-time pricing, and incentive-based DR (IBDR) schemes, e.g., direct load control, interruptible service and demand bidding. DR programs can be deployed in the industrial, commercial and residential sector. This thesis focuses on IBDR in the residential sector. Details on existing DR schemes are presented in section 2.1.

1.1. Motivation for residential IBDR

Although some IBDR programs have penalties for not responding to DR requests, many programs are completely voluntary. Participants choose to receive incentives for their DR resources. PBDR programs on the other hand, obligate consumers to reduce loads during peak hours or they face high electricity prices. Therefore, IBDR programs are considered reward-wise programs, whereas PBDR programs are considered punishment-wise programs. In general, the voluntary nature of reward-wise programs makes people more positive and responsive in the long term, whereas the obligatory nature of punishment-wise programs makes people nervous and responses are more transient [4].

IBDR programs deployed in the U.S. contribute 93% of load reductions compared to PBDR programs [10]. The major share of these reductions come from the industrial and commercial sector. However, IBDR still has untapped potential in the residential sector, which makes up the largest share of total energy demand [36]. Moreover, residential loads can provide a more reliable, flexible and continuous response than large industrial loads [3].

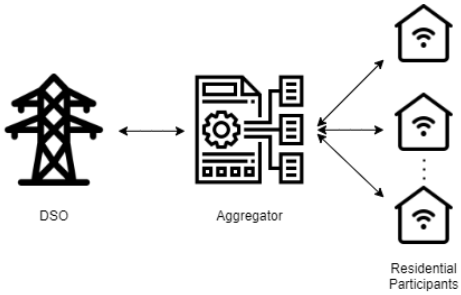


Figure 1.1: The aggregator combines DR resources and makes them available to utilities.

Despite the aforementioned advantages, in practice, residential consumers hardly participate in IBDR programs. IBDR schemes are less suitable for residential consumers for several reasons. The following challenges describe the complications of residential participation and how they can be addressed:

1. Residential loads are relatively small compared to industrial loads. In the case of interruptible service programs it is not viable for utilities to establish contracts with residential consumers for minor reductions. Similarly, residential bids in demand bidding programs are too small for utilities to accept them. Typically, utilities set minimum requirements for levels of active load which residential consumers often do not meet. To solve this issue, residential loads can be combined by an aggregator. Aggregators are a key stakeholder in the electricity market acting as an intermediary between utilities and end consumers [48]. By aggregating residential loads they meet the minimum requirements and aggregators make residential DR resources available to utilities (Fig. 1.1).

2. Prediction of grid congestion is complicated and a DR requests are often submitted in short notice. Effective response to these requests requires either direct control of residential appliances by the utility or immediate action by residents, which is infeasible in practice. A Home Energy Management System (HEMS) is a device installed in households to allow residents to efficiently manage energy consumption. The HEMS is connected to smart meters and smart plugs in order to access consumption measurements. Moreover, the HEMS is able to locally control connected smart appliances such as the dishwasher and the washing machine. The HEMS enables automated response to DR requests in real time without requiring explicit action from the residents.

3. For interruptible service programs utilities require detailed information on consumption behavior of the participant to establish an agreement for predefined load reduction. Residential consumers have higher privacy requirements than industrial consumers which may complicate this. To preserve the privacy of residential consumers, no other parties should have access to preferences of the residents. Similarly, direct load control may be invasive to participant privacy and affect their comfort.

Many studies have proposed approaches to automate residential DR programs. Centralized approaches require exhaustive information about individual participants, which may not be available or may cause privacy issues [21][44]. In addition, these centralized approaches rely on conventional optimization methods like linear programming [16][31] or dynamic programming [46], which make real-time computation infeasible for large number of participants in the program.

Reinforcement Learning (RL) is a promising approach for decision-making in IBDR programs for several reasons. First, RL methods require no initial knowledge of the problem. RL agents learn through interaction with the environment and therefore are independent of the organization of the program and do not require information about other participants. Second, Multi-agent Reinforcement Learning (MARL) controls multiple agents in a decentralized fashion, which allows scaling up the number of participants. Third, once trained, RL agents can decide nearly instantly, facilitating future real-time control applications. RL problems are formulated as a Markov Decision Process (MDP), a framework for sequential decision-making, where a decision in one time step influences the next. Details on RL and MDPs are presented in Section 2.2.

1.2. Related Work

Currently, a great deal of research is devoted to the application of RL to DR [38]. Some studies focus on the industrial sector. [30] 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. [17] 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.

Prior works on RL for DR in the residential sector mainly focus on home energy management in a single household. [41] present an RL-based approach to DR for a single residential household or a 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. [23] 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 Deep Q-network based on convolutional networks learns to schedule the load blocks.

Others focus on DR on the scale of the wholesale electricity market. [47] propose a voluntary incentive-based DR program targeting retail consumers with smart meters paying a flat electricity price. Load serving entities provide consumers with 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. [44] 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 market players.

However, the previous approaches rely on centralized algorithms instead of MARL. The following works propose decentralized MARL methods for load scheduling of appliances in a collection of households. [12] propose a framework for scheduling the consumption profile of appliances in multiple households modeled 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. [45] apply a cooperative RL approach to schedule controllable appliances of multiple households such that utility costs are minimized. 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. [20] proposes a real-time RL algorithm for incentive-based DR programs that supports service providers (aggregators) to purchase energy flexibility as a DR 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 profit and dissatisfaction are balanced in the objective of the service provider. [40] propose a similar method that includes PV generation and [42] 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 propose single-agent RL focused on either the home energy management perspective or on the aggregator perspective. To the best of the author's knowledge, no approach has been proposed that applies MARL to residential IBDR for the benefit of both aggregators and consumers. A comparative overview is given in Table 1.1.

1.3. Research Questions and Contributions

To address the identified research gap, this thesis aims to answer the following research question: *How can RL induce flexibility in residential demands to relieve grid congestion?*

The research question is split up in two sub-questions:

	Residential	**IBDR**	**MARL**
[30][17]		✓	✓
[41][23]	✓		
[47][44]	✓	✓	
[12][45]	✓		✓
[20][40][42]	✓	✓	

Table 1.1: Overview of related work and the aspects they consider.

1. How can a residential IBDR program be designed to relieve grid congestion?
2. How can RL automate a residential IBDR program while preserving privacy and considering heterogeneous preferences of residential consumers?

To answer the research questions this thesis proposes a novel decentralized Multi-Agent Reinforcement Learning approach for Incentive-based DR (MARL-iDR). The proposed approach considers simultaneously a single Aggregator Agent (AA) and multiple participant agents aiming to maximize their individual rewards. The aggregator learns to deploy a suitable incentive based on total electricity demand and 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, dishwasher and air conditioning, while minimizing dissatisfaction of the residents. The optimal power assignment is achieved through the proposed internal optimization of a Disjunctively Constrained Knapsack Problem (DCKP) [6].

The main contributions of this thesis are:

1. An environment that formulates an IBDR program including an aggregator and multiple residential participants as an MDP. The overall objective in the IBDR program is to maintain residential demands below capacity limits determined by the DSO, providing it with flexibility to relieve grid congestion. The environment model internally solves the DCKP to minimize participant dissatisfaction, taking the participant demand as input and schedules household appliances as output. The environment model aims to answer sub-question 1.
2. A MARL algorithm suitable for automating residential IBDR using Deep Q-Learning. The algorithm makes real-time decisions in a decentralized fashion for the aggregator and its residential participants, while preserving participants' privacy and accounting for heterogeneous preferences. The MARL algorithm aims to answer sub-question 2.

1.4. Thesis Outline

The rest of this thesis is organized as follows. Section 2 provides background information on DR and RL. Section 3 presents the proposed MARL-iDR approach. In Section 3.1, an IBDR program is formulated as an MDP and Section 3.2 describes the MARL algorithm. In Section 3.2, the proposed MARL-iDR algorithm is described. In Section 4, the results of a case study are presented to evaluate the performance of the approach. Finally, Section 5 concludes the thesis and presents recommendations for future research.

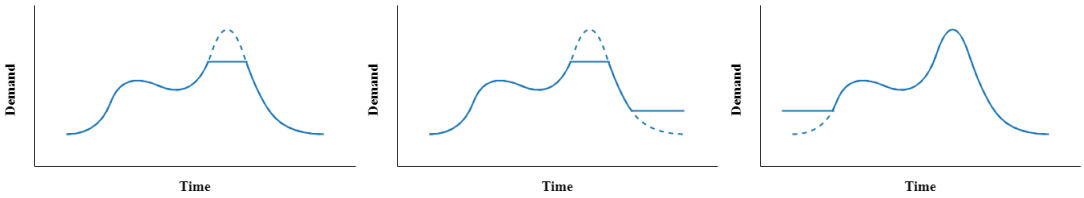
2

Theoretical Background

2.1. Demand Response

Demand Response (DR) is defined as a change in consumption patterns of electricity consumers in response to financial incentives in times when the network reliability is at risk [15]. DR is beneficial to multiple stakeholders in the electricity market, including the Transmission System Operator (TSO), DSO, and end consumers [5]. The TSO is responsible for managing and transmitting high-voltage electricity from generation sources to sub-stations. DR enables the TSO to improve the reliability of the high-voltage grid. Having flexible demand at critical moments, e.g., failure of a generation source or a power line, reduces the risks of power shortages and power outages. The DSO can use DR for maintaining capacity limits of the low-voltage grid at distribution level. DR reduces the risk of congestion in power lines or sub-stations. End consumers benefit from DR by having lower energy bills or financial rewards for participation.

An indirect method of DR is strategic conservation, where the overall load is reduced. Trivial approaches to strategic conservation are improving energy efficiency (e.g., replacing traditional appliances with energy-efficient appliances) and energy saving (e.g., turning off lights when leaving the room). A more direct method of DR is load management. Load management distinguishes three categories: peak clipping, load shifting and valley filling. Peak clipping is the reduction of the peak load to avoid exceeding the capacity of the system (Fig. 2.1a). Reducing load is achieved by curtailing the power consumption of specific appliances, e.g., dim lights or turn down the AC. Load shifting is moving demand from peak hours to off-peak hours (Fig. 2.1b). This is achieved by scheduling operation of appliances with a specific run time to a later moment. Finally, valley filling is the increase of load during times of oversupply for stability in the system (Fig. 2.1c). Appliances are intentionally turned on to increase demand, typically without additional costs.



(a) Peak clipping (b) Load shifting (c) Valley filling

Figure 2.1: Categories of load management.

Generally, electricity consumers are classified into four sectors: industrial, commercial, residential and transport (Fig. 2.2). Utilities offer DR programs in the former three. Existing DR programs can be categorized into two approaches: price-based and incentive-based. In general, PBDR is suitable for the residential sector and IBDR is suitable for the industrial and commercial sectors [19].

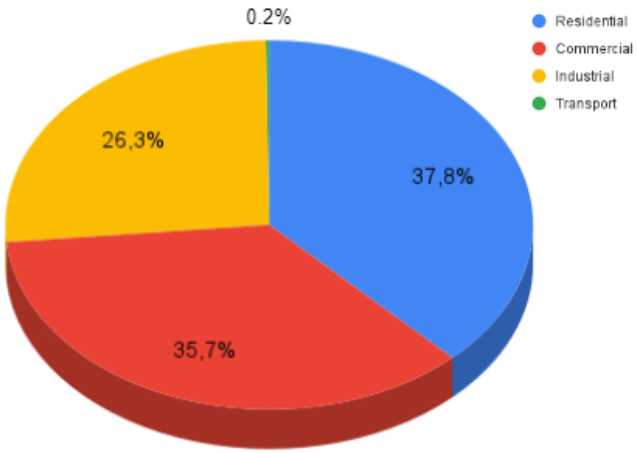


Figure 2.2: Share of consumption on the U.S. electricity market by sector in 2020 [36].

In price-based DR (PBDR), participants are offered varying electricity prices depending on the value and cost of electricity in that period. When demand exceeds the supply, electricity prices will be high. On the other hand, when more electricity is produced then consumed the prices will be low. The electricity price encourages consumers to individually manage their consumption by reducing it during peak hours or by shifting it to less congested hours. PBDR methods are mostly suitable for residential consumers [19]. PBDR schemes include:

- **Time of Use**: Consumers are offered two price rates, a low price during off-peak hours and a high price during on-peak hours. Usually, the prices and hours are predefined, e.g., a day tariff of 15 cents per kilowatt hour between 7:00 and 0:00 and a night tariff of 6 cents per kilowatt hour between 0:00 and 7:00. Time of Use is a common pricing scheme and is offered in countries worldwide [18].

- **Critical Peak Pricing**: During short periods with a critical peak, the price goes up according to the peak. Utilities can announce the critical peak in fixed periods or in variable periods based on predefined criteria. Several major utilities in the US offer this pricing scheme.

- **Real Time Pricing**: Electricity prices are adapted to demand and production costs during the whole day, typically in hourly intervals. There are two main challenges to the application of this scheme. Firstly, it relies on continuous real-time data exchange, which is not favorable for consumers. Secondly, the large-scale data processing increases the complexity of the whole system. Several experimental Real Time Pricing schemes in the US show significant reductions of peak demand and flattening of consumption [37].

In incentive-based DR (IBDR), utilities submit a DR request to encourage participants to lower or shift consumption at critical moments in return for financial incentives. When participants respond to a request, the load reduction is considered a DR resource and participants receive payment. In IBDR programs, typically a contractual agreement between the utility and participant is established [19]. IBDR schemes include:

- **Direct load control**: According to a detailed agreement between consumer and utilities, utilities can remotely control some consumer loads. Notices for operation are announced a short time ahead. Participants require remote control equipment, allowing utilities to turn on, turn off, reschedule, or curtail electrical appliances [29].

- **Interruptible service**: At risk of grid congestion, participants are requested to reduce loads to an agreed level. Failing to respond leads to penalties. The frequency and the duration of DR requests are limited [1].

- **Demand bidding**: Utilities announce the total amount of electricity that must be curtailed. Consumers can bid for the amount on the basis of their own situation and the wholesale market. The announcement is normally released one day ahead [2].

2.2. Reinforcement Learning

Reinforcement Learning (RL) is a field of research in machine learning that has gained much interest over the years. Inspired by neuroscience and human psychology, RL is a computational approach to goal-directed learning from interaction with the natural environment [35]. A learning agent learns a mapping from situations, i.e., states of its environment, to actions in order to maximize its reward. The learner does not know which actions to take, but must instead discover which actions yield the most reward by trying them. It should be able to find a balance between exploiting actions that have been rewarding in the past and exploring undiscovered actions. In some cases, rewards are not received immediately, but require a sequence of decisions that may be individually unprofitable. Therefore, an ideal learner does not only consider immediate rewards, but also rewards in the future.

2.2.1. Markov Decision Process

The RL learning problem is typically formulated as a Markov Decision Process (MDP), which is a framework for sequential decision-making. 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 (not on previous states). In an MDP, the agent and the environment interact in discrete time steps $t = 0, 1, 2, \ldots$. The agent observes the state of the environment, decides upon an action, receives a reward and transit to the next time step with corresponding state as shown in Fig. 2.3. In the case of an finite horizon MDP, the number of time steps T is finite and the sequence of all steps 0 to T is called an episode.

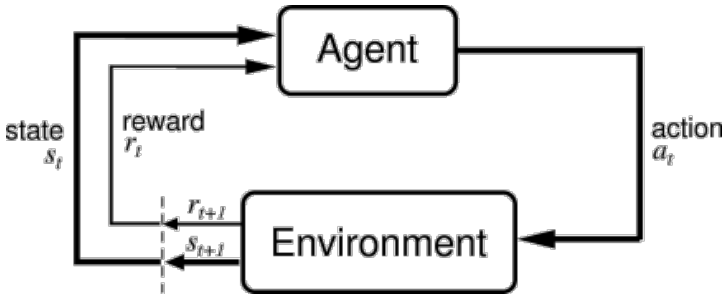


Figure 2.3: Interaction between the agent and the environment.

An MDP is formally defined as a tuple $\langle S, A, p, r \rangle$

- S is the set of possible states of the environment called the state space.
- A is the set of valid actions called the action space.
- p is the transition probability function, i.e., probability that the environment will transit to the next state s' given the current state s and an action a.
- r is the reward function, i.e., the numerical reward after taking action a in state s and transitioning to state s'.

The task of the agent is to maximize the accumulated reward in the long term rather than maximize the immediate reward. More formally, the objective of the agent is to find a policy π that maximizes the expected discounted return G over finite time horizon T in each time step t:

$$G_t = \mathbb{E}\left\{ \sum_{k=0}^{T} \gamma^k r_{t+k+1} \right\} \tag{2.1}$$

where $\gamma \in [0,1]$ is the discount rate that determines the value of future rewards relative to immediate rewards. The finite-horizon MDP assumes there always is a finite state, possibly after a fixed number of steps.

Most RL approaches involve estimating the optimal state-action value functions. The optimal value of a state-action pair satisfies the Bellman optimality equation:

$$V^*(s,a) = \sum_{s' \in S} p(s,a,s') \left[r(s,a,s') + \gamma \max_{u'} V^*(x',u') \right] \tag{2.2}$$

If the optimal value of each state-action pair is available, the optimal policy is computed by selecting for state each s the action that has the largest value:

$$\pi^*(s) = arg\max_{a} V^*(s,a) \tag{2.3}$$

2.2.2. Q-learning

A widely used RL algorithm is Q-learning [39]. Q-learning is an approach to approximate the values of state-action pairs. Q-learning turns Eq. 2.2 into an iterative approximation procedure:

$$Q_{t+1}(s_t,a_t) = Q_t(s_t,a_t) + \eta \left[r_{t+1} + \gamma \max_{a'} Q_t(s_{t+1},a') - Q_t(s_t,a_t) \right] \tag{2.4}$$

A Q-learning agent initializes a Q-table with all valid state-action pairs with arbitrary Q-values. In each time step t the agent observes the transition to time step $t+1$ and the corresponding reward $(s_t, a_t, s_{t+1}, r_{t+1})$ and uses Eq. (2.4) to update its Q-values. The term between the square brackets is the temporal difference, i.e., the difference between the current estimate of the state-action value and the updated estimate in the next time step. Learning rate η controls the size of the update step.

2.2.3. Action Selection

The agent should be able to find a balance between exploration, i.e., discovering new, potentially profitable actions and exploitation, i.e., selecting the best action known so far. The most used action selection strategy is called ϵ-greedy. The agent selects a random action with probability $\epsilon \in [0,1]$ and the greedy action (best action) with probability $(1-\epsilon)$. In the variant of ϵ-greedy with decay, ϵ decreases over time with decay rate $\delta \in [0,1]$ with a minimum of ϵ_{min}:

$$\epsilon_t = min(\epsilon_{max}, \epsilon_{t-1} \cdot \delta) \tag{2.5}$$

With this strategy, the agent benefits from extensive exploration early in training and refinement of the policy in later stages [9].

2.2.4. Deep Q-Learning

The Q-learning algorithm described in Section 2.2.2 is very effective and is a popular choice for developers of RL algorithms because of its simplicity. However, if the amount of state-action pairs is significantly large or continuous, Q-learning becomes infeasible for two reasons: (1) the memory required to store Q-values in the Q-table increases directly with the amount of state-action pairs and (2) the time required to explore every state-action pair grows rapidly.

Deep Reinforcement Learning (DRL) combines the field of Deep Learning with RL. The idea of DRL algorithms is to incorporate deep learning approaches in RL algorithms, typically by representing the policy as a

neural network. Deep Q-Learning (DQL) is a DRL variant of Q-learning [25]. The concept of DQL is to replace the Q-table with a neural network, i.e., a Deep Q-Network (DQN) or policy network. Instead of directly storing state-action pairs in a Q-table, the Q-values are approximated by the DQN that takes a state as input and the Q-values of the actions in response to that state as output.

However, there is a challenge when we compare DQL to conventional deep learning. In deep learning, neural networks are trained given an input and a target that matches the input. With DQL, that target is represented by the temporal difference from Eq. (2.4): $r_{t+1} + \gamma \max_{a'} Q_t(s_{t+1}, a')$. The issue is that this target is non-stationary, i.e., the target is constantly changing, making training of the network unstable. To fixate the target a second DQN can be used to predict the target. This so called target network θ^T is not trained, hence target predictions are stable. Instead, the parameters of the target network are replaced every N^t time steps. A second challenge is that neural networks are usually trained on batches of training data to stabilize training. In RL however, agents only observe the current state and reward. To solve this issue, Experience Replay is applied. Every experience $(s_t, a_t, s_{t+1}, r_{t+1})$ is stored in a Replay Buffer B. The policy network can now be trained on a batch of experiences sampled from B. The main reason for training on random batches is to break correlations between consecutive samples which make training inefficient. The procedure for RL training with DQNs is presented in Algorithm 1.

Algorithm 1 DQN training procedure

Initialize policy network θ with arbitrary parameters
Initialize target network $\theta^T = \theta$
Initialize replay buffer B,
Initialize $\epsilon_0 \leftarrow 1.0$
Initialize buffer size N^b, sample batch size N^s and target replace interval N^t
for all episodes **do**
    Update $\epsilon_t = \epsilon_{t-1} \cdot \delta$
    for all time steps t **do**
        $x \leftarrow$ random number $\in [0, 1]$
        if $x \leq \epsilon$ **then**
            Select random action $a_t \in A$
        else
            $a_t \leftarrow \arg\max_a Q(s_t, a|\theta)$
        end if
        Execute action a_t and observe reward r_t and state s_{t+1}
        Store experience s_t, a_t, r_t, s_{t+1} in B
        if $|B| \geq N^b$ **then**
            Remove oldest experience tuple from B
        end if
        Sample batch of N^s experiences from B
        for all experiences (s_t, a_t, r_t, s_{t+1}) **do**
            $y = r_t + \gamma \max_{a'} Q(s_{t+1}, a'|\theta^T)$
            Perform gradient descent step on θ with loss $|| y - Q(s_t, a_t|\theta) ||^2$
        end for
        if time step t equals interval N^t **then**
            Replace parameters of θ^T with parameters of θ
        end if
    end for
end for

2.2.5. Multi-Agent Reinforcement Learning

MDPs can be generalized to the multi-agent case, called a stochastic game. A stochastic game is formulated as a tuple $\langle S, A_1, \ldots, A_n p, r_1, \ldots, r_n \rangle$ where n is the number of agents. The state space S and the transition probability function p are the same as in the single-agent case, but each agent i has its own set of valid

actions A_i and its own reward function r_i. The rewards depend on the combination of actions by all agents. In the case that the reward function is equal for all agents ($r_1 = \cdots = r_n$), the problem is fully cooperative and in the case that the reward functions oppose each other (in the two-agent case $R_1 = -R_2$), the problem is fully competitive. Reward functions may be different for some or all agents, in which case the stochastic game is considered a mixed game [8].

Multi-agent Reinforcement Learning (MARL) decentralizes the problem by dividing tasks to different agents. If many different decisions need to be made in each time step, single-agent RL may become infeasible because of the huge joint action space. However, due to this decentralization non-stationarity arises since the agents train simultaneously. Instead of only interacting with the stationary environment, i.e., the response of the environment (reward) remains the same during training, the agent also interacts with other agents. Since agents continuously update their policies, the rewards change during training even though the state and the action of the agent is the same. In other words, the optimal policy for an agent changes as the other agents change their policy. Although developers of MARL algorithms often include some form of coordination between agents, fully independent training has also proven effective.

3

Proposed Approach: MARL-iDR

In this section, the proposed approach, MARL-iDR, is described in detail. MARL-iDR is a Multi-agent RL method for automating responses in an incentive-based DR program. The approach consists of an environment model of an IBDR program, described in Section 3.1 and a MARL algorithm, described in Section 3.2, applied to the environment model.

3.1. Environment Model

This section formulates an IBDR program as the environment of a multi-agent MDP. The architecture for the program is shown in Fig. 3.1 and its components are described in detail throughout this section. The overall model considers multiple agents: one for the aggregator (AA) and one for each participant (PAs). The AA distributes incentives to the different PAs where each represents a residential household.

3.1.1. Assumptions for IBDR 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 similar to the interruptible service program in [26]. The aggregator incentivizes its enlisted participants to realize the target reduction through peak clipping and load shifting, i.e., lowering and delaying consumption of household appliances respectively. Valley filling is not considered in the program.

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. The estimated demand is then used as the CBL. 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 approach to demand estimation, the current day is matched with ten previous similar days and the average consumption is computed considering changes in weather conditions. Details for computation of the CBL are found in [32].

3.1.2. Participant Agent

Each PA represents a household in the IBDR program. The set of all PAs is $\mathscr{H}$. Each PA $i \in \mathscr{H}$ can control a set of appliances $\mathscr{D}_i$. In practice, the PA would be integrated in the 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 shifting appliances. The environment model for each household is defined by its

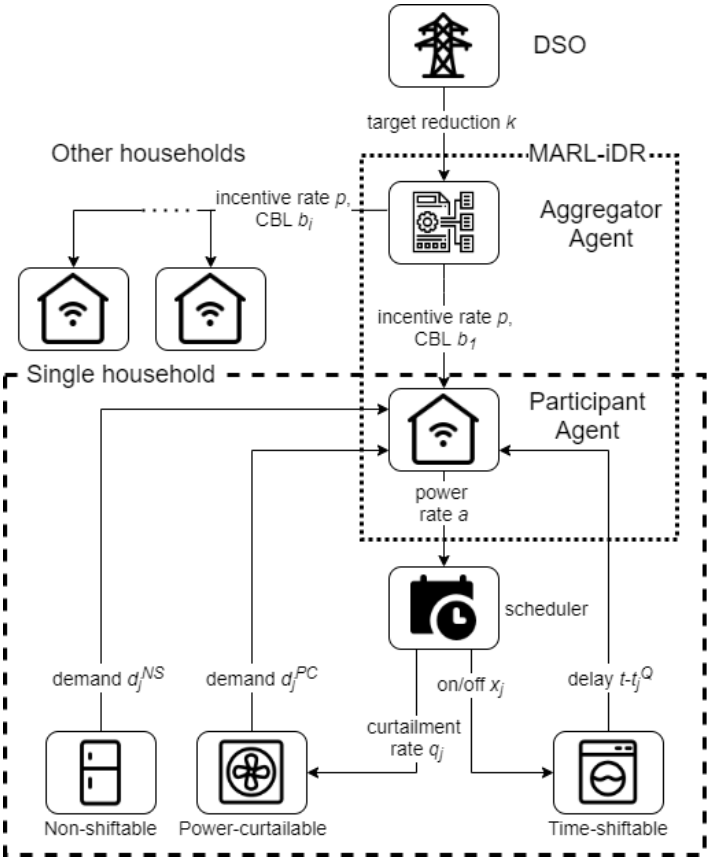


Figure 3.1: Architecture of the environment

state, actions, rewards and the scheduler.

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

- Time-shiftable appliances are either on or off. It is assumed that they consume a constant amount of power. 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). Time-shiftable appliances are useful for load management in the form of load shifting (Fig. 2.1b). Their state is determined by the difference between the current time step t and the initial request time step t^I_j, i.e., the current delay $t - t^I_j$.
- Power-curtailable appliances allow to lower power consumption, but do not allow delay of usage (e.g., changing the setpoint of an AC, dimming lighting). Power-curtailable appliances are useful for load management in the form of peak clipping (Fig. 2.1a). Their state is the variable power demand in kilowatts d^{PC}_j.
- Non-shiftable appliances must run at all times without delay or curtailment. These appliance share a single state defined as the total non-shiftable power demand in kilowatts d^{NS}_j

The state $s_{t,i}$ of a household i at time step t combines the states of all appliances $j \in \mathscr{D}_i$.

In addition, the PA observes the incentive rate p_t passed from the AA and its own projected CBL $b_{t,i}$. The incentive rate determines the amount of money received in cents per kilowatt reduction, where the reduction

is measured against the CBL. For example, if the incentive is 5 cents per kilowatt reduction, the CBL is 3 kilowatt and the consumption is 1 kilowatt, the PA will receive 10 cents in that time slot.

Finally, the observation of PA i in time step t is $o_{t,i}^{PA} = \{s_{t,i}, b_{t,i}, p_t\}$. For the remainder of Section 3.1.2 and Section 3.1.3, the subscripts t and i are dropped as all equations apply to time step t and household i.

Action This thesis 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 3.1.3 matches the appliances to the limit $l = a \cdot d$ where d is the total demand of time-shiftable and power-curtailable appliances. The resulting total power consumption e 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 shifting and curtailing appliances. 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. Here, b and e are in kilowatts and p in cents per kilowatt. The financial reward u paid from AA to PA is

$$u = p \cdot \max(0, b - e), \tag{3.1}$$

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 loose anything.

Second, shifting or curtailing 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:

$$c_j^{TS} = \beta_j \left(t + 1 - t_j^I\right)^2 \quad \forall j \in TS, \tag{3.2}$$

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 [43]. In the case of power-curtailable appliance $j \in PC$, the dissatisfaction cost is a convex function of the power curtailment.

$$c_j^{PC} = \beta_j \left(\frac{1}{m} \cdot q_j \cdot d_j\right)^2 \quad \forall j \in PC, \tag{3.3}$$

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 [14]. β_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 preference or that can be learned through feedback of the residents.

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

$$r^{PA} = u - \sum_{j \in \mathcal{D}} c_j \tag{3.4}$$

3.1.3. Scheduler

The scheduler is a part of household i and would be integrated in the HEMS in practice. The scheduler determines the optimal assignment of power to appliances in TS and PC based on demand limit l. The scheduler

is a combinatorial optimization formulated as a DCKP [6]:

$$\text{minimize} \quad \sum_{j \in PC} c_j^{PC} + \sum_{j \in TS} (1 - x_j) \cdot c_j^{TS} \tag{3.5}$$

$$\text{subject to} \quad \sum_{j \in PC} \frac{1}{m} \cdot q_j \cdot d_j + \sum_{j \in TS} x_j \cdot d_j \leq l \tag{3.6}$$

$$x \in \{0,1\},\ q \in \{0,1,\ldots,m\} \tag{3.7}$$

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. (3.2) and Eq. (3.3). After solving the DCKP, the overall power demand is

$$e = \sum_{j \in PC} \frac{1}{m} \cdot q_j \cdot d_j + \sum_{j \in TS} x_j \cdot d_j \tag{3.8}$$

3.1.4. Aggregator Agent

The objective of the aggregator is to reduce the aggregated power consumption below a certain limit in kilowatts as per agreement with the DSO with a minimal amount of incentive paid to the participants.

State The observation of the AA in time step t is $o_t^{AA} = \{d_t, k\}$, where d_t is the aggregated demand of all households $i \in \mathcal{H}$ in kilowatt and k is the target reduction set by the DSO in kilowatt.

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

Reward The reward of the AA is

$$r_t^{AA} = -\Big(\rho \cdot e_t^+ + (1 - \rho) \cdot \sum_{i \in \mathcal{H}} u_{t,i}\Big), \tag{3.9}$$

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 payed to the PAs as defined in Eq. 3.1. The trade-off between the two terms is determined by weighting factor ρ. 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.

3.2. MARL Algorithm

The proposed MARL method is a multi-agent DQL algorithm where the AA and PA have indirectly opposing reward functions, i.e., actions in favor of the AA may have negative influence on the reward of the PA and vice versa. All agents are trained simultaneously, hence, the agents deal with a moving target where the optimal policy changes as opposing agents change their policies. Simultaneous learning leads to non-stationary problems which invalidate most theoretical guarantees of single-agent RL, e.g., the guarantee of convergence [8]. Despite these limitations, simultaneous learning has found numerous applications because of its simplicity [13][22].

MARL-iDR effectively trades-off the exploration with exploitation. MARL-iDR uses the action selection strategy of ϵ-greedy with decay, i.e., with probability ϵ the agent selects a random action where ϵ decreases over time with decay rate δ [9]. With this strategy, MARL-iDR benefits from extensive exploration early in training and refinement of the policy in later stages.

The training procedure for MARL-iDR is presented in Algorithm 2. 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. Each episode 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 receives 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 π:

$$\pi(s) = \underset{a}{argmax}\, Q(s, a|\theta) \tag{3.10}$$

Algorithm 2 MARL-iDR training procedure

Initialize $\theta^{AA}, \theta^{T,AA}, B^{AA}$
Initialize $\theta_i^{PA}, \theta_i^{T,PA}, B_i^{PA} \quad \forall i \in \mathscr{H}$
Initialize $\epsilon_0 \leftarrow 1.0$
Initialize $\delta, \quad 0 < \delta < 1$
for all episodes **do**
    Initialize target reduction k
    Initialize rewards $r_0^{AA}, r_0^{PA} \leftarrow 0$
    $\epsilon_t = \epsilon_{t-1} \cdot \delta$
    for all time steps t **do**
        Observe demand d_t and calculate CBLs b_t
        $o_t^{AA} \leftarrow \langle d_t, k \rangle$
        Add $\langle o_{t-1}^{AA}, p_{t-1}, r_{t-1}^{AA}, o_t^{AA} \rangle$ to B
        Train θ^{AA} given B
        Select p_t using ϵ-greedy given θ^{AA}
        for all PAs $i \in \mathscr{H}$ **do**
            Observe state of appliances $s_{t,i}$, CBL $b_{t,i}$ and incentive rate p_t
            Compute dissatisfaction costs $c_{t,i}$ given $s_{t,i}$
            $o_{t,i}^{PA} \leftarrow \langle s_{t,i}, b_{t,i}, p_t \rangle$
            Add $\langle o_{t-1,i}^{PA}, a_{t-1,i}, r_{t-1,i}^{PA}, o_t^{PA} \rangle$ to B
            Train θ_i^{PA} given B
            Select $a_{t,i}$ using ϵ-greedy given θ_i^{PA}
            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

4

Case Study

A case study is performed to test the effectiveness of the environment model and the MARL algorithm. The case is a small neighborhood in Austin, Texas, US, during the summer months when air conditioning consumes significant amounts of power due to warm weather conditions. All households have an HEMS installed that has access to power consumption measurements of smart appliances and can manage them automatically. The MARL-iDR approach is evaluated taking five aspects into consideration: 1. Convergence of training curves 2. Policies learned by the agents 3. Information exchange and privacy preservation 4. Computational efficiency and 5. Economics for the aggregator considering a varying weighting factor ρ. The approach is compared to the original demand, i.e., when no demand response is performed, and to a centralized baseline approach that computes the optimal myopic actions.

4.1. Simulation Data and Test Setup

4.1.1. Dataset Processing

This case study uses appliance requests generated from electricity data from a PecanStreet dataset [27]. The dataset is a 15-minute interval time series for 25 real households in Austin, Texas, US. The dataset is loaded using Pandas [24]. Each entry in the data frame contains the timestamp, the household ID and average real power consumption in kilowatts over the 15-minute interval for several household appliances and groups of power outlets. For the case study, 5 major appliances are selected based on two criteria. First, the appliance must have a large power consumption such that rescheduling or curtailing has a significant effect on the demand. Second, the appliance must be owned and used by most of the households. The EV, washing machine, dryer, dishwasher and AC fit these criteria. All other appliances in the dataset are considered non-shiftable appliances.

An important assumption in the environment model is that the power consumption of time-shiftable appliances is constant while running its program. In practice, however, power consumption of most appliances fluctuates. Time-shiftable appliances are considered running when the real average power consumption exceeds 0.1 kilowatts. The average power consumption for each time-shiftable appliance is taken from [28].

Data for several timestamps and households is missing or unreliable. For missing entries, the real average power consumption is assumed 0 kilowatts. The unreliable entries are negative values and unrealistically large values. Therefore, values of real average power consumption of power-curtailable appliances and non-shiftable appliances are clipped between 0 kilowatts and a maximum amount of kilowatts.

Households have diverse preferences regarding management of their appliances. For example, some households prefer the washing machine to finish its program before the dishwasher, while others prefer the opposite. To introduce heterogeneous preferences, the dissatisfaction coefficients β for the appliances are sam-

Appliance	**Type**	**Demand (kW)**	**Dissatisfaction coefficient, 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 4.1: Household appliances and their parameters, e.g, non-interruptible (*NI*) and interruptible (*I*).

pled from a normal distribution. The type, demand and dissatisfaction coefficients of the appliances selected for the simulations are in Table 4.1. The exact coefficients of this case study and the impact of the coefficients on management of the appliances are presented in Appendix A.3.

After processing the PecanStreet data, the load curves are obtained. The load curve is the total continuous electricity demand of a household over time in kilowatt. Fig. 4.1 presents the load curve of each individual household between 12:00 PM and 0.00 AM on July 1. The load curves are extremely diverse due to the heterogeneous usage of household appliances. The load curves provide complex and realistic inputs for the environment model.

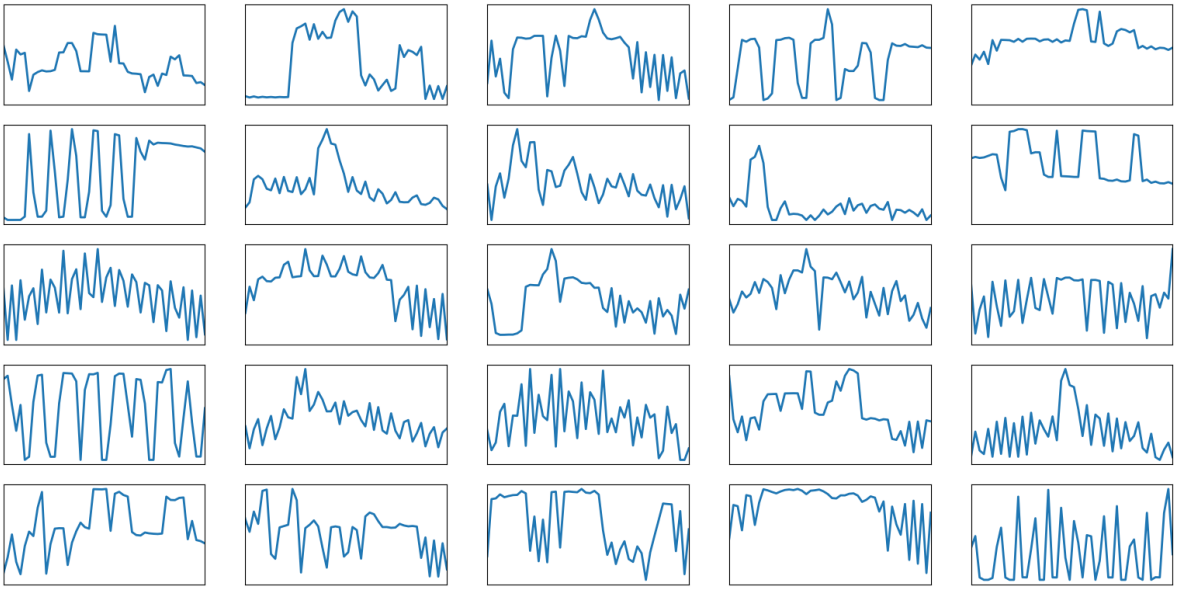


Figure 4.1: Load curves of the individual households.

4.1.2. Implementation Details

The AA and the PAs each have an individual DQN implemented with Keras [11]. Each DQN is a feedforward neural network. The network for the AA consists of one input layer of 2 input neurons corresponding to its observation space, two hidden layers of 32 neurons and one output layer of 11 output neurons corresponding to its action space. Similarly, the network for each PA consists of one input layer of 8 input neurons corresponding to its observation space, two hidden layers of 32 hidden neurons and one output layer of 11 output neurons corresponding to its action space. The outputs of each layer are activated with a ReLU activation function. Finally, the weights are initialized by sampling from a normal distribution and the biases are initialized as 0.

The MARL-iDR algorithm is trained for 5000 episodes, discretized in $T = 96$ time steps of 15 minutes and ran-

domly sampled from the training period April 1, 2018, to October 31, 2018 (excluding the validation period July 1, 2018 to July 31, 2018). The action space for the PAs and the AAs is 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 learning rate $\eta = 0.001$, the discount rate $\gamma = 0.9$ and ϵ-decay rate $\delta = 0.999$. Weighting factor ρ is 0.5. Finally, the target reduction k is defined at 80% of the peak demand. The 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.

4.1.3. Myopic Baseline

A baseline is used to compare the performance of the proposed MARL algorithm. 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 $p^* = argmax\{r^{AA}|a_i^*, \forall i \in H\}$. Algorithm 3 presents the procedure for calculating the optimal responses for both the AA and all PAs. The procedure is called every time step. Since the optimal response of the AA depends on the response of the PAs, first the optimal response of each PA is calculated (line 3) and saved for each incentive rate in optimal power rate matrix M. Subsequently, the optimal response of the AA is calculated given M (line 16) and saved as p^*. Unlike RL methodology, this myopic baseline requires full knowledge about the model, e.g., the reward function, and is only able to consider immediate rewards (short-sighted).

Algorithm 3 Myopic Baseline

```
1: Initialize optimal power rate matrix M
2: Initialize optimal incentive rate p*
3: for all incentives p ∈ A^AA do
4:     for all PA i ∈ ℋ do
5:         Initialize optimal reward r*
6:         for all power rate a ∈ A^PA do
7:             Calculate PA reward r given p and choosing a
8:             if r > r* then
9:                 r* ← r
10:                M_{p,i} ← a
11:            end if
12:        end for
13:    end for
14: end for
15: Initialize optimal reward r*
16: for all incentives p ∈ A^AA do
17:    Calculate PA reward r given optimal responses M_p
18:    if r > r* then
19:        r* ← r
20:        p* ← p
21:    end if
22: end for
23: Return AA action p* and PA action vector M_{p*}
```

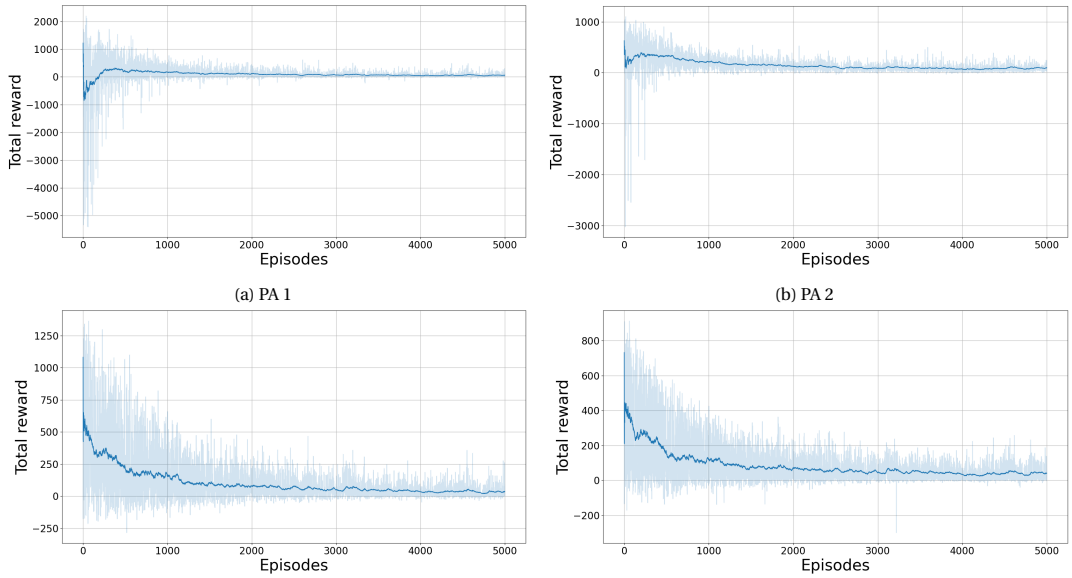
4.2. Results

In this section, results of the case study are presented and discussed. Some results are explained through illustration on a single day and for a single household. It should be mentioned that this day and household are selected to best illustrate a pattern in the results, although all days and households contribute to the overall performance of the approach.

4.2.1. Training Curves

Fig. 4.2 and Fig. 4.3 present the training curves of four arbitrary PAs and the AA respectively. The y-axis shows the cumulative reward over the course of a single episode. The dark-blue line shows the exponential weighted average with a window size of 100 to show how the average total reward converges over time. For all training curves, it can be observed that the rewards spread over a wide range, even at later stages of training, due to the diverse nature of each episode, i.e., each training day has different potential for reward. Nevertheless, the training curves for both the AA and the PAs are converging to a stable policy. The other PAs show similar training curves and converge as well.

Regarding the training curve of PA 1 and PA 2, the cumulative reward grows quickly up to episode 500, after which it starts to decrease. For PA 3 and PA 4, the training curves are decreasing from the first episode. This can be explained due to the fact that the AA is constantly changing its policy as well, causing the PA to earn less incentive in later stages.



(a) PA 1

(b) PA 2

(c) PA 3

(d) PA 4

Figure 4.2: Learning curves for four PAs.

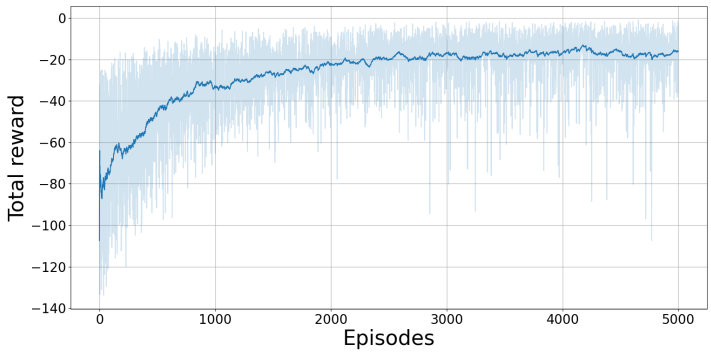


Figure 4.3: Learning curve for the AA.

4.2.2. Load Reductions and Incentive Rates

MARL-iDR reduces loads during peak hours. The results are presented in Table 4.2. 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 the peak load and PAR even further, slightly exceeding the target

	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 4.2: Results averaged per day in July.

reduction with a total of 0.49 kWh, whereas this "surplus consumption" for MARL-iDR is 3.93 kWh. MARL-iDR sometimes results in a second peak that exceeds the target reduction k. This unwanted effect of load shifting to a second peak is known as rebound effect [7]. Fig. 4.4 illustrates this behavior, showing the impact of MARL-iDR on the load curve on July 1st. The aggregated load (Fig. 4.4b) is unchanged before 14:30. During hours where the original load exceeds the target reduction, i.e., peak hours, 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 (90.7 kW) is lower than the first, original peak (102.7 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). In comparison, 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. 4.4c 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. The error of the CBL is analyzed in Appendix A.1.

4.2.3. Dissatisfaction Costs and Appliance Scheduling

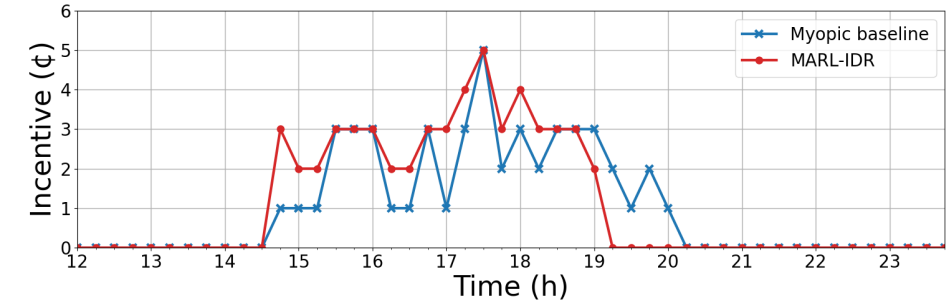
The appliance schedule of the household from Fig. 4.4c is presented in Fig. 4.5. Fig. 4.5a shows for each appliance the originally requested time step and the actual scheduled time step. The washing machine and the EV delay for as long as incentives are offered. As soon as the incentive rate drops to 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. 4.5b shows the trade-off between incentives gained and dissatisfaction caused by rescheduling appliances. The dissatisfaction from delaying the EV and the washing machine is increasing until scheduled at 19.15. A detailed analysis of the appliance scheduling optimization is presented in Appendix A.2.

4.2.4. Preserving Privacy

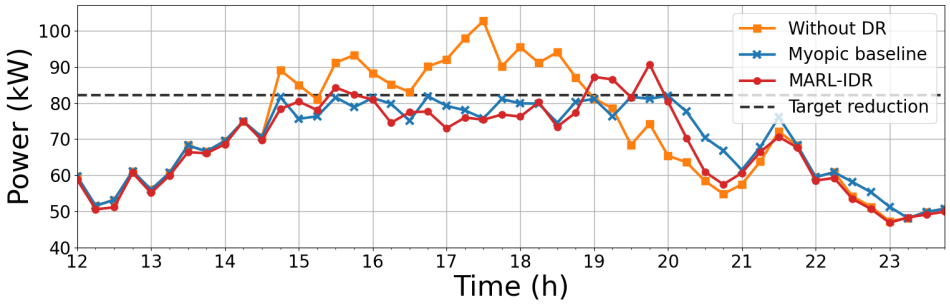
The proposed MARL-iDR preserves privacy which is legally required. MARL-iDR outperforms any centralized scheduler at the AA (e.g., myopic baseline) as they require exact information about the reward function to calculate the best responses. In practice, this means the aggregator must know participants'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

4.2.5. Economic Analysis

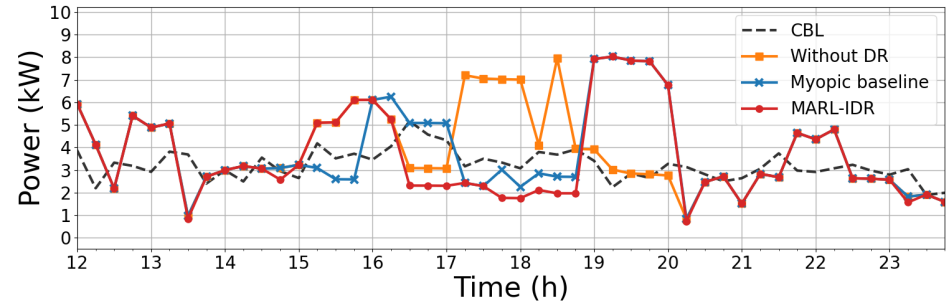
MARL-iDR trains local agents balancing the economics between the aggregator and all participants. The key metric in this balance is the incentive rate. The incentive rates selected by the AA are shown in Fig. 4.4a. The shape of this rate curve matches the original aggregated demand curve in Fig. 4.4b. MARL-iDR stops offering incentives after 19.00 and the myopic baseline stops after 20.00. The myopic baseline offers less or equal



(a) Incentive rates selected by the AA



(b) Aggregated load curve.



(c) Load curve of an individual household

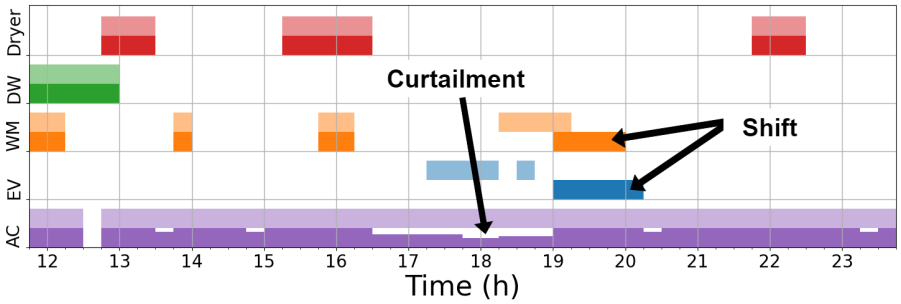
Figure 4.4: Load reductions and incentive rates for MARL-iDR compared to the myopic baseline and the case without DR.

incentives optimizing the target reduction between 14.45 and 18.15 which results in lower financial rewards for the PAs.

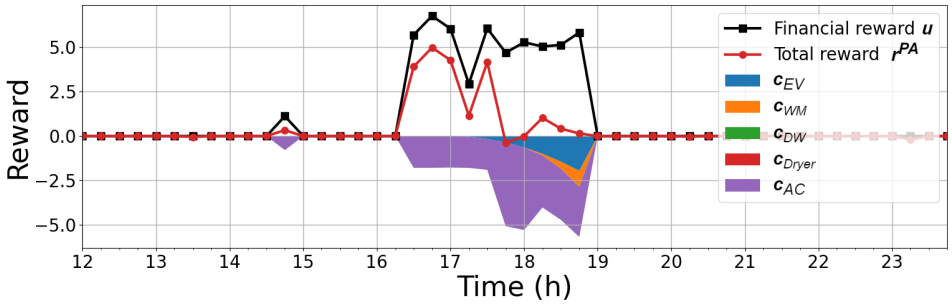
The design of the program is important for its success toward a fair balance in financial costs and gains for all parties. The design (and the balance) is controlled at the AA through selecting the weighting factor ρ in Eq. (3.9). The selection of ρ ultimately determines the trade-off between financial costs and violation of the target reduction. A study on this parameter ρ is in Fig. 4.6. The larger ρ, the larger the penalty on surplus consumption and the smaller the incentive cost. For small values ρ, there is a significant amount of consumption exceeding the target while only a little incentive is paid to participants. On the other hand, when ρ is large, the aggregator tries to push the surplus consumption down to 0 by offering increasing amounts of incentive.

4.2.6. 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 residential IBDR. The computation times of scheduling appliances of a single household in a single time step are compared with the baseline. The myopic baseline is a centralized approach computing all possibilities before the PA can decide the optimal scheduling, resulting in a computational complexity of $\mathcal{O}\left(|A^{AA}||\mathcal{H}||A^{PA}|\right)$. The used implementation took on average 1.86 s, and for 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 ap-



(a) Household schedule with requests (light blocks) and assignments by PA (dark blocks). The 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.

Figure 4.5: Appliance scheduling and dissatisfaction.

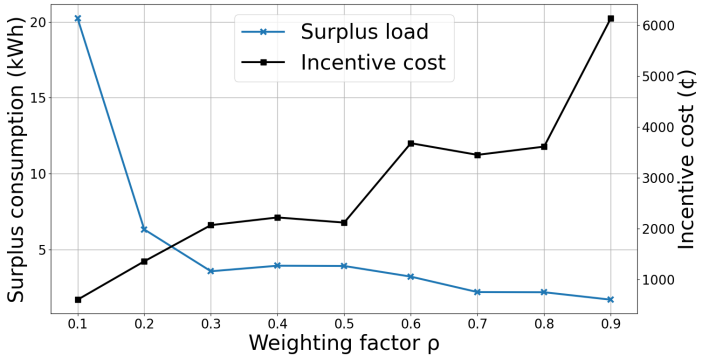


Figure 4.6: The aggregator balances financial cost and surplus consumption with ρ.

pliances independently, the actual schedule can be computed in 2 ms. As 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, and the key advantage is that almost unlimited many PAs can be considered at the same time. The time of training MARL-iDR with one AA and 25 PAs for 5000 episodes is ~ 12 h which only has to be done once before deployment.

4.3. Discussion

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 financially 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 due to 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 including 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). Curtailing the AC provides a considerable amount of load reductions in these months. However, the effectiveness has not been tested in winter months or at different locations. AC demands are likely to be much smaller in January or in colder cities. In these scenarios, it makes more sense to curtail (water) heating appliances or lighting.

5

Conclusion and Recommendations

5.1. Conclusion

This thesis proposes MARL-iDR, a novel decentralized MARL approach, for a residential incentive-based DR program formulated as an MDP. The approach balances interests of the aggregator and multiple residential participants while preserving the privacy of participants and making decisions in real-time. The aggregator agent learns a policy that dispatches suitable incentives to participants based on total energy demand and a target reduction, while minimizing financial costs. The participant agent learns to respond to these incentives by reducing consumption to a fraction of the original demand. The participant agents curtail or shift requested household appliances based on the selected consumption reduction using a novel Disjunctively Constrained Knapsack Problem optimization, while minimizing residents' dissatisfaction. The approach demonstrates the ability to induce flexibility in residential demands to relieve grid congestion. A case study shows a PAR reduction of 9% compared to the original demand.

The first research question of this thesis was: *How can a residential IBDR program be designed to relieve grid congestion?* This question is answered by environment model of an IBDR program suitable for deployment in the residential sector. The IBDR program aims to provide DSOs with demand-side flexibility at critical moments. The DSO determines a target reduction necessary to relieve congestion and the participants provide the DR resources through aggregation to achieve the target reduction.

The second research question of this thesis was: *How can RL automate a residential IBDR program while preserving privacy and considering heterogeneous preferences of residential consumers?* This question is answered by the proposed MARL algorithm. The algorithm automates responses of households to DR requests in real-time. The algorithm minimizes dissatisfaction of residents with diverse consumption patterns and heterogeneous preferences. Finally, the algorithm controls multiple agents, allowing it to make decisions in a decentralized fashion with minimal information exchange to preserve the privacy of residents.

5.2. Recommendations

This work delivers initial steps towards a proof of concept and towards a possible future real-time application. Further research is needed before integrating RL agents with the HEMS in households. As a recommendation, the following points require further research to improve the performance of the MARL-iDR approach:

- In the current approach, all RL agents are acting independently without any awareness of other agents or any form of coordination. Demand-side flexibility could be increased when agents start coordinating their strategies. For example, PAs could coordinate scheduling of appliances to prevent the rebound effect. Such coordination requires some information exchange which may violate privacy requirements. A more implicit approach is to shape the reward function of PAs to include a collective goal, implic-

itly compelling agents to collaborate to maximize rewards. For example, the agents receive a bonus if the collectively reach the target reduction or, on the opposite, they do not receive any incentives if the target reduction is not met.

- Despite the fact that the agents are training in a non-stationary environment, the agents converge to a stable policy. It is recommended to study what the impact of this non-stationarity is and if providing a stable environment improves the final policy. For example, if the policy of all PAs is fixed, will exclusive training of the AA lead to convergence to a better policy?
- The environment is modelled with time steps of 15 minutes. Since MARL-iDR promises real-time responses, the approach should allow deployment at smaller time intervals, e.g., seconds or even milliseconds. It is hypothesized that by correctly tuning the dissatisfaction coefficients the approach can make decisions at any time interval larger than the time required to make decision. Empirical evidence is required to confirm the hypothesis that the approach generalizes over different time step sizes.

As a recommendation, the following points require further research to bring the approach one step closer to practical implementation:

- Little is known about the generalizability of the approach. The case study was performed on a limited amount of households and a limited amount of appliances. The case was set in Austin, Texas, US, with warm weather conditions. To allow a widespread application, it is recommended to investigate the performance under different conditions, e.g., the number of participants, more diverse sets of household appliances, different locations, weather conditions and periods of the year.
- As in interruptible service programs, the target reduction is pre-defined by utilities. In the case study of this work the target reduction is set to 80% of the peak demand. For practical implementation, it is useful to utilities to have a variable target to provide the flexibility they need to prevent grid congestion. It is recommended to test a range of targets and find the feasible limits of this target.
- For simplicity, the time-shiftable appliances are assumed to operate with constant power consumption. In practice, the power consumption of household appliances fluctuates. It requires further investigation to test the effectiveness with real power consumption data as input.
- Today, a large portion of residential houses include Photo-Voltaic (PV) systems as well as Energy Storage Systems (ESS). This allows storage of locally generated solar energy and consumption during peak hours. A more realistic case study includes PV systems and ESS in households.

Bibliography

[1] H. A. Aalami, M. Parsa Moghaddam, and G. R. Yousefi. "Demand response modeling considering Interruptible/Curtailable loads and capacity market programs". In: *Applied Energy* 87.1 (Jan. 2010), pp. 243–250. ISSN: 03062619. DOI: 10.1016/j.apenergy.2009.05.041.

[2] Christopher O. Adika and Lingfeng Wang. "Demand-side bidding strategy for residential energy management in a smart grid environment". In: *IEEE Transactions on Smart Grid* 5.4 (2014), pp. 1724–1733. ISSN: 19493053. DOI: 10.1109/TSG.2014.2303096.

[3] Ailin Asadinejad et al. "Evaluation of residential customer elasticity for incentive based demand response programs". In: *Electric Power Systems Research* 158 (May 2018), pp. 26–36. ISSN: 03787796. DOI: 10.1016/j.epsr.2017.12.017.

[4] P. Teimourzadeh Baboli et al. "Customer behavior based demand response model". In: *IEEE Power and Energy Society General Meeting* (2012). DOI: 10.1109/PESGM.2012.6345101.

[5] V. S.K.Murthy Balijepalli et al. "Review of demand response under smart grid paradigm". In: *IEEE PES International Conference on Innovative Smart Grid Technologies* (2011), pp. 236–243. DOI: 10.1109/ISET-INDIA.2011.6145388.

[6] Mariem Ben Salem et al. "Optimization algorithms for the disjunctively constrained knapsack problem". In: *Soft Computing* 22.6 (Dec. 2016), pp. 2025–2043. ISSN: 1433-7479. DOI: 10.1007/S00500-016-2465-7. URL: https://link-springer-com.tudelft.idm.oclc.org/article/10.1007/s00500-016-2465-7.

[7] Zane Broka and Karlis Baltputnis. "Handling of the Rebound Effect in Independent Aggregator Framework". In: *International Conference on the European Energy Market, EEM* (Sept. 2020). DOI: 10.1109/EEM49802.2020.9221943.

[8] Lucian Busoniu, Robert Babuška, and Bart De Schutter. "Multi-agent reinforcement learning: An overview". In: *Studies in Computational Intelligence* 310 (2010), pp. 183–221. DOI: 10.1007/978-3-642-14435-6{_}7.

[9] Olivier Caelen and Gianluca Bontempi. "Improving the Exploration Strategy in Bandit Algorithms". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 5313 LNCS (2007), pp. 56–68. DOI: 10.1007/978-3-540-92695-5{_}5.

[10] Peter Cappers, Charles Goldman, and David Kathan. "Demand response in U.S. electricity markets: Empirical evidence". In: *Energy* 35.4 (Apr. 2010), pp. 1526–1535. ISSN: 03605442. DOI: 10.1016/j.energy.2009.06.029.

[11] Chollet, Francois, and et. al. *Keras*. 2015. URL: https://keras.io.

[12] Hwei-Ming Chung et al. "Distributed Deep Reinforcement Learning for Intelligent Load Scheduling in Residential Smart Grids". In: (June 2020).

[13] Robert Crites and Andrew Barto. "Improving Elevator Performance Using Reinforcement Learning". In: *Advances in Neural Information Processing Systems* 8 (1995).

[14] Murat Fahrioglu and Fernando L. Alvarado. "Using utility information to calibrate customer demand management behavior models". In: *IEEE Transactions on Power Systems* 16.2 (May 2001), pp. 317–323. DOI: 10.1109/59.918305.

[15] Federal Energy Regulatory Commission (FERC). "Assessment of Demand Response and Advanced Metering Staff Report". In: (2008). URL: https://www.ferc.gov/industries-data/electric/power-sales-and-markets/demand-response/reports-demand-response-and.

[16] Armen Gholian, Hamed Mohsenian-Rad, and Yingbo Hua. "Optimal Industrial Load Control in Smart Grid". In: *IEEE Transactions on Smart Grid* 7.5 (Sept. 2016), pp. 2305–2316. DOI: 10.1109/TSG.2015.2468577.

[17] Xuefei Huang et al. "Demand Response Management for Industrial Facilities: A Deep Reinforcement Learning Approach". In: *IEEE Access* 7 (2019), pp. 82194–82205. DOI: 10.1109/ACCESS.2019.2924030.

[18] International Renewable Energy Agency (IRENA). "Innovation landscape brief: Time-of-use tariffs". In: (2019). URL: www.irena.org.

[19] 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., Jan. 2017, pp. 203–230. ISBN: 9780081012628. DOI: 10.1016/B978-0-08-101753-1.00007-3.

[20] Renzhi Lu and Seung Ho Hong. "Incentive-based demand response for smart grid with reinforcement learning and deep neural network". In: *Applied Energy* 236 (Feb. 2019), pp. 937–949. ISSN: 03062619. DOI: 10.1016/j.apenergy.2018.12.061.

[21] Sabita Maharjan et al. "Dependable demand response management in the smart grid: A stackelberg game approach". In: *IEEE Transactions on Smart Grid* 4.1 (2013), pp. 120–132. DOI: 10.1109/TSG.2012.2223766.

[22] Maja J. Matarić. "Learning in multi-robot systems". In: *Lecture Notes in Computer Science* 1042 (1995), pp. 152–163. DOI: 10.1007/3-540-60923-7{_}25.

[23] Alwyn Mathew, Abhijit Roy, and Jimson Mathew. *Intelligent Residential Energy Management System using Deep Reinforcement Learning*. Tech. rep. 2020.

[24] Wes McKinney. "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference* (2010), pp. 56–61. DOI: 10.25080/MAJORA-92BF1922-00A.

[25] Volodymyr Mnih et al. "Playing Atari with Deep Reinforcement Learning". In: (Dec. 2013). URL: https://arxiv.org/abs/1312.5602v1.

[26] Pacific Gas and Electric Company (PG%E). *Base Interruptible Program (BIP)*. URL: https://www.pge.com/en_US/large-business/save-energy-and-money/energy-management-programs/demand-response-programs/base-inerruptible/base-inerruptible.page?WT.mc_id=Vanity_bip.

[27] *Pecan Street Inc.* URL: https://www.pecanstreet.org/.

[28] *Power Consumption of Typical Household Appliances*. URL: https://www.daftlogic.com/.

[29] Badri Ramanathan and Vijay Vittal. "A framework for evaluation of advanced direct load control with minimum disruption". In: *IEEE Transactions on Power Systems* 23.4 (2008), pp. 1681–1688. ISSN: 08858950. DOI: 10.1109/TPWRS.2008.2004732.

[30] Martin Roesch et al. "Smart grid for industry using multi-agent reinforcement learning". In: *Applied Sciences (Switzerland)* 10.19 (Oct. 2020), pp. 1–20. ISSN: 20763417. DOI: 10.3390/app10196900.

[31] Prabodi Ruwanthika Senevirathne et al. "Optimal Residential Load Scheduling in Dynamic Tariff Environment". In: *ICIIS - Proceedings* (Dec. 2019), pp. 547–552. DOI: 10.1109/ICIIS47346.2019.9063296.

[32] R. Sharifi, S. H. Fathi, and V. Vahidinasab. "Customer baseline load models for residential sector in a smart-grid environment". In: *Energy Reports* 2 (Nov. 2016), pp. 74–81. ISSN: 2352-4847. DOI: 10.1016/J.EGYR.2016.04.003.

[33] Benjamin K Sovacool et al. "The Power Production Paradox: Revealing the Socio-Technical Impediments to Distributed Generation Technologies". In: (Apr. 2006). URL: https://vtechworks.lib.vt.edu/handle/10919/27058.

[34] Anna Stawska et al. "Demand response: For congestion management or for grid balancing?" In: *Energy Policy* 148 (Jan. 2021). ISSN: 03014215. DOI: 10.1016/j.enpol.2020.111920.

[35] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction Second edition*. The MIT Press, 2018.

[36] U.S. Energy Information Administration. *Annual Energy Outlook 2021*. URL: https://www.eia.gov/outlooks/aeo/.

[37] John S. Vardakas, Nizar Zorba, and Christos V. Verikoukis. “A Survey on Demand Response Programs in Smart Grids: Pricing Methods and Optimization Algorithms”. In: *IEEE Communications Surveys and Tutorials* 17.1 (Jan. 2015), pp. 152–178. ISSN: 1553877X. DOI: 10.1109/COMST.2014.2341586.

[38] José R. Vázquez-Canteli and Zoltán Nagy. “Reinforcement learning for demand response: A review of algorithms and modeling techniques”. In: *Applied Energy* 235 (Feb. 2019), pp. 1072–1089. ISSN: 0306-2619. DOI: 10.1016/J.APENERGY.2018.11.002.

[39] Christopher J. C. H. Watkins and Peter Dayan. “Q-learning”. In: *Machine Learning 1992 8:3* 8.3 (May 1992), pp. 279–292. ISSN: 1573-0565. DOI: 10.1007/BF00992698. URL: https://link.springer.com/article/10.1007/BF00992698.

[40] Lulu Wen et al. “Modified deep learning and reinforcement learning for an incentive-based demand response model”. In: *Energy* 205 (Aug. 2020), p. 118019. ISSN: 03605442. DOI: 10.1016/j.energy.2020.118019.

[41] Zheng Wen, Daniel O’Neill, and Hamid Reza Maei. “Optimal Demand Response Using Device Based Reinforcement Learning”. In: *IEEE Transactions on Smart Grid* 6.5 (Jan. 2014), pp. 2312–2324. URL: https://arxiv.org/abs/1401.1549v2.

[42] Hongsheng Xu et al. “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, Sept. 2020. ISBN: 9781728157481. DOI: 10.1109/APPEEC48164.2020.9220364.

[43] Xu Xu et al. “A Multi-Agent Reinforcement Learning-Based Data-Driven Method for Home Energy Management”. In: *IEEE Transactions on Smart Grid* 11.4 (July 2020), pp. 3201–3211. ISSN: 19493061. DOI: 10.1109/TSG.2020.2971427.

[44] Mengmeng Yu and Seung Ho Hong. “Incentive-based demand response considering hierarchical electricity market: A Stackelberg game approach”. In: *Applied Energy* 203 (Oct. 2017), pp. 267–279. ISSN: 03062619. DOI: 10.1016/j.apenergy.2017.06.010.

[45] Chi Zhang et al. “A cooperative multi-agent deep reinforcement learning framework for real-time residential load scheduling”. In: *IoTDI 2019 - Proceedings of the 2019 Internet of Things Design and Implementation* (Apr. 2019), pp. 59–69. DOI: 10.1145/3302505.3310069.

[46] Ying Jun Angela Zhang et al. “Profit-maximizing planning and control of battery energy storage systems for primary frequency control”. In: *IEEE Transactions on Smart Grid* 9.2 (2018), pp. 712–723. DOI: 10.1109/TSG.2016.2562672.

[47] Haiwang Zhong, Le Xie, and Qing Xia. “Coupon incentive-based demand response: Theory and case study”. In: *IEEE Transactions on Power Systems* 28.2 (2013), pp. 1266–1276. ISSN: 08858950. DOI: 10.1109/TPWRS.2012.2218665.

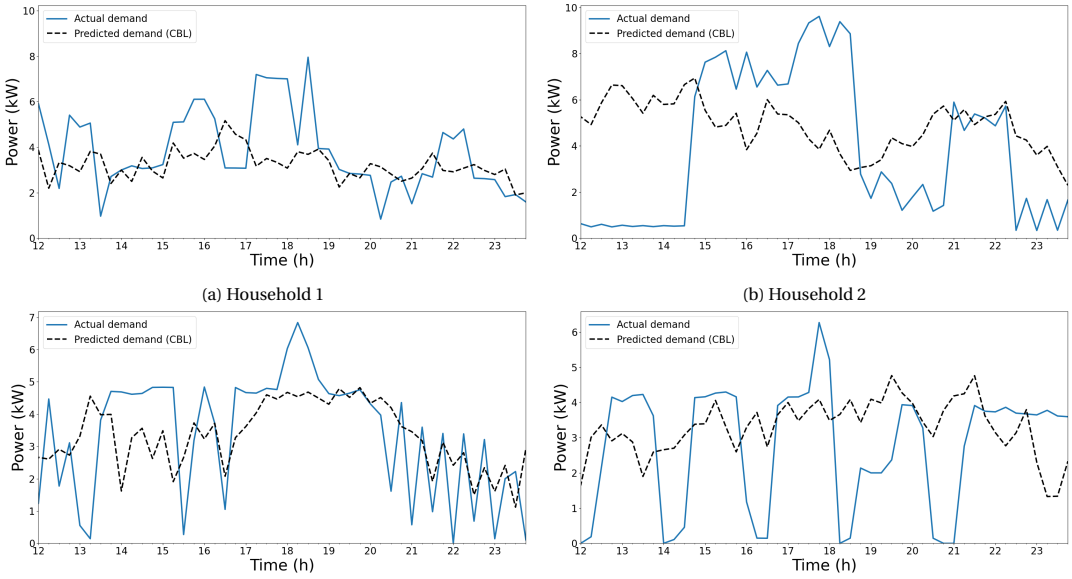
[48] Kaile Zhou and Shanlin Yang. “Smart Energy Management”. In: *Comprehensive Energy Systems*. Vol. 5-5. Elsevier, Jan. 2018, pp. 423–456. DOI: 10.1016/B978-0-12-809597-3.00525-3.

A

In-depth Analysis

A.1. Customer Baseline Load (CBL)

In IBDR programs, demand reductions are measured against a reference demand, i.e., CBL. The exact demand of participants in future time steps is unknown, hence, aggregators have to estimate the demand of their participants. In practice, utilities deploy different methods in existing IBDR programs to determine the CBL. Most of these methods rely on historical data and statistical analysis. Fig. A.1 shows the predicted demand or CBL against the actual demand for the first four households from Fig. 4.1. For households 1 and 3, the patterns in the load curves are similar to historical patterns, hence it can be seen that the CBL roughly corresponds to the actual demand, except for some outliers, e.g., the peak between 17:00 and 18:00 of household 1. Often such peaks deviate from the CBL when appliances are requested outside the regular pattern of the household. Household 2 and 4 on the other hand, show less correspondence between the actual demand and CBL. This is explained since there are less patterns in the consumption behavior of these households. Mismatches in actual demand and CBL affect behavior of the PA. When the CBL is higher than the demand, e.g., between 12:00 and 14:30 in household 2, the PA will receive incentives even without reducing consumption. When the CBL is lower than the demand, e.g., between 15:00 and 18:30 in household 2, the PA may decide that it can not reduce demands below the CBL and chooses maximum consumption to avoid dissatisfaction costs. In the illustrative example in Appendix A.2 the demand is peaking above the CBL, however the PA still reduces the demand significantly to reach below the CBL.



(a) Household 1

(b) Household 2

(c) Household 3

(d) Household 4

Figure A.1: Actual demand and predicted demand (CBL) for four households.

A.2. Appliance Scheduling

This thesis proposes an action space of discrete power rates for the PAs to control the power demand of the household. The selected action is followed by an appliance scheduling optimization that converts a power rate into a power assignment to household appliances. This appendix analyzes this transition from power rate to power assignment in depth.

Time step $t = 73$ (18.15) is considered on the same day and the same PA as the case study (Fig. A.2). In this specific time step, residents request operation of multiple household appliances simultaneously which contributes to the complexity of the power assignment. The EV has been delayed for 4 time steps up to this point and which causes a dissatisfaction of 1 for denying the requests. A new request is submitted to run the washing machine (WM), which causes a smaller dissatisfaction of 0.1 for denying the request. In addition, the AC is demanding a significant amount of power which can be curtailed. The EV demands a constant 4 kilowatt, the WM demands 1 kilowatt, the AC demands 2.47 kilowatt and an additional 0.64 kilowatt is demanded by non-shiftable appliances, 8.11 kilowatt in total. The AA offers the PAs 3 cents per kilowatt reduction and has determined a CBL of 3.80 kilowatt as a reference consumption for this particular PA. Table A.1 shows the how the appliance scheduling optimizer assigns power based on the power rate selected by the PA.

If power rate $a = 1.0$ is selected, all appliances receive the requested power. In this case, no dissatisfaction is caused, but also no power is reduced. Therefore, the CBL is exceeded and no incentive is earned. If the power rate is smaller, the scheduler first denies the WM power ($a = 0.9, a = 0.8$), then denies the EV power ($a = 0.7, a = 0.6, a = 0.5$). The AC is only slightly curtailed at $a = 0.8$. However, delaying either of them is not a significant power reduction to consume less than the CBL. For $a = 0.4$ both the EV and the WM are delayed and the AC and the non-shiftable appliances consume a total of 3.11 kilowatt, subceeding the CBL and earning incentives. For power rates lower than 0.4 power consumption of the AC is curtailed. In the extreme case of $a = 0.0$, no power is assigned at all, causing significant dissatisfaction, mostly for the AC (18.26). The optimal myopic action in this time step is $a = 0.3$ (1.55), the PA selected a near optimal power rate of $a = 0.2$ (1.03). However, optimal myopic action may not give the highest long-term reward.

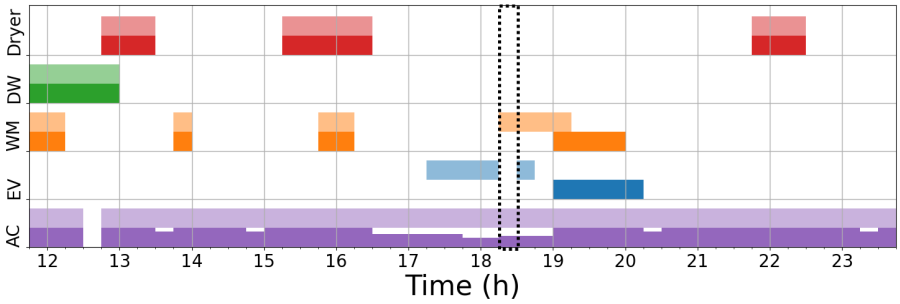


Figure A.2: Time step selected for analysis of appliance scheduling.

	Power rate	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Consumption	**WM**	0	0	0	0	0	1	1	1	0	0	1
	EV	0	0	0	0	0	0	0	0	4	4	4
	AC	0	0.74	1.48	2.22	2.47	2.47	2.47	2.47	1.97	2.47	2.47
	Total	0.64	1.38	2.12	2.86	3.11	4.11	4.11	4.11	6.61	7.11	8.11
Dissatisfaction	**WM**	0.1	0.1	0.1	0.1	0.1	0	0	0	0.1	0.1	0
	EV	1	1	1	1	1	1	1	1	0	0	0
	AC	18.26	8.95	2.92	0.18	0	0	0	0	0.73	0	0
	Total	19.36	10.05	4.02	1.28	1.1	1	1	1	0.83	0.1	0
Reduction		3.16	2.42	1.68	0.94	0.70	0	0	0	0	0	0
Incentives earned		9.49	7.27	5.05	2.83	2.09	0	0	0	0	0	0
Reward		-9.87	-2.78	1.03	1.55	0.99	-1	-1	-1	-0.83	-0.1	0

Table A.1: The conversion from power rate to power assignment by the appliance scheduling optimization for time step is $t = 73$ on July 1st. The residents request the washing machine, EV and AC. For each power rate also the reduction relative to the CBL (3.80), the earned incentive and the reward are given.

A.3. Impact of Dissatisfaction Coefficients

The dissatisfaction coefficient β controls for each appliance how fast the dissatisfaction of the residents builds up when it is delayed or curtailed. From Eq. (3.2) and Eq. (3.3), it can be derived that dissatisfaction increases quadratically as a function of the delay and the power curtailment respectively. To confirm the impact of the dissatisfaction coefficients on the dissatisfaction, the average curtailment for the AC and the average delay of the EV are investigated under different values of the dissatisfaction coefficient. The training setup is the same as the case study described in Section 4 except all PAs are trained independently with different values for β. To make sure the different coefficients do not affect the AA, its policy is fixed, i.e., the policy the AA converged to in the case study.

Fig. A.3 shows how the dissatisfaction coefficient of the AC affects power curtailment. The non-zero AC power curtailments for each household and in each time step over the month July are concatenated and the boxplot is presented. The median is slightly decreasing from 0.36 kilowatt for $\beta = 1.0$ to 0.18 kilowatt for $\beta = 5.0$. The impact is also clearly observed in the range in which power curtailments fall. For $\beta = 1.0$, power curtailment reaches up to 3.3 kilowatt. For $\beta = 4.0$, the maximum power curtailment is below 1.2 kilowatt, although for $\beta = 5.0$ several outliers have larger power curtailments.

Fig. A.4 shows how the dissatisfaction coefficient of the EV affects the delay of charging the EV. The non-zero delays for each EV request from each household over the month July are concatenated and the boxplot is presented. It can be observed that the median decreases when the coefficient is increased. For $\beta = 0.01$, the median of delay is 2.25 hours in contrast to a median delay of 1 hour for $\beta = 0.07$. The same trend is observed in the maximum delays. For $\beta = 0.01$, the maximum delay over all EV requests is 6.75 hours. For $\beta = 0.07$, the maximum delay is just 2.75 hours.

In practice, the dissatisfaction coefficient could be an internal parameter in any HEMS that can be explicitly controlled by the residents or learned by the HEMS through feedback from the user about their level of

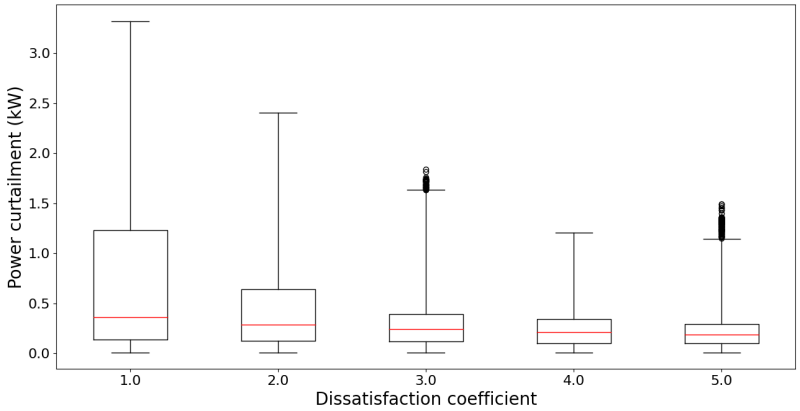


Figure A.3: Effect of the dissatisfaction coefficient on the power curtailment of the AC.

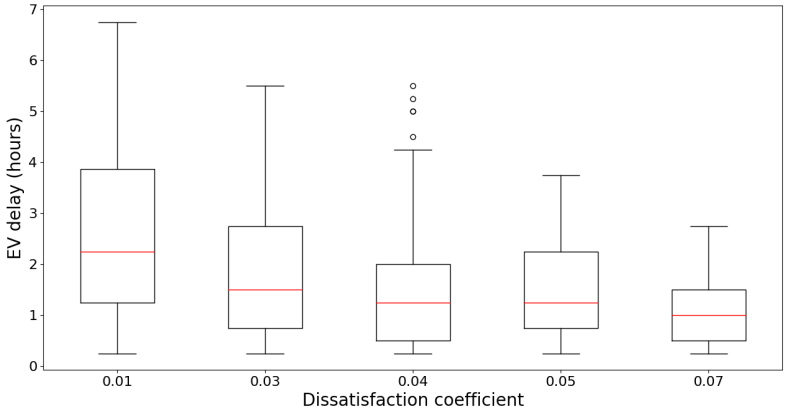


Figure A.4: Effect of the dissatisfaction coefficient on the delay of charging the EV.

dissatisfaction.

Varying dissatisfaction coefficients represent heterogeneous household preferences. Table A.2 provides the exact values for the dissatisfaction coefficients for each household and each appliance as used in the case study.

ID	AC	EV	WM	DW	Dryer
1	3.000	0.040	0.100	0.060	0.200
2	2.843	0.023	0.038	0.071	0.079
3	2.009	0.090	0.257	0.010	0.493
4	3.131	0.022	0.103	0.082	0.151
5	2.302	0.028	0.010	0.027	0.294
6	3.088	0.010	0.168	0.017	0.552
7	2.363	0.010	0.222	0.062	0.361
8	1.918	0.082	0.137	0.010	0.010
9	2.932	0.078	0.066	0.032	0.391
10	2.248	0.032	0.010	0.017	0.283
11	3.006	0.031	0.059	0.010	0.177
12	2.596	0.070	0.185	0.127	0.148
13	5.322	0.091	0.192	0.067	0.142
14	1.584	0.112	0.010	0.010	0.127
15	2.536	0.062	0.127	0.039	0.415
16	2.462	0.104	0.200	0.010	0.183
17	2.339	0.036	0.010	0.010	0.156
18	2.358	0.010	0.067	0.034	0.084
19	4.472	0.010	0.133	0.010	0.174
20	4.514	0.067	0.031	0.010	0.098
21	2.751	0.027	0.176	0.010	0.155
22	1.045	0.042	0.087	0.029	0.378
23	3.828	0.012	0.105	0.131	0.025
24	2.841	0.019	0.217	0.010	0.010
25	2.232	0.010	0.133	0.112	0.260

Table A.2: Dissatisfaction coefficients used in the case study.