Multi-agent reinforcement learning for the coordination of residential energy flexibility

Flora Valérie Jeanne Charbonnier

Pembroke College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2023

Abstract

This thesis investigates whether residential energy flexibility can be coordinated without sharing personal data at scale to achieve a positive impact on energy users and the grid.

To tackle climate change, energy uses are being electrified at pace, just as electricity is increasingly provided by non-dispatchable renewable energy sources. These shifts increase the requirements for demand-side flexibility. Despite the potential of residential energy to provide such flexibility, it has remained largely untapped due to cost, social acceptance, and technical barriers. This thesis investigates the use of multi-agent reinforcement learning to overcome these challenges.

This thesis presents a novel testing environment, which models electric vehicles, space heating, and flexible household loads in a distribution network. Additionally, a generative adversarial network-based data generator is developed to obtain realistic training and testing data. Experiments conducted in this environment showed that standard independent learners fail to coordinate in the partially observable stochastic environment. To address this, additional coordination mechanisms are proposed for agents to practise coordination in a centralised simulated rehearsal, ahead of fully decentralised implementation.

Two such coordination mechanisms are proposed: optimisation-informed independent learning, and a centralised but factored critic network. In the former, agents lean from *omniscient* convex optimisation results ahead of fully decentralised coordination. This enables cooperation at scale where standard independent learners under partial observability could not be coordinated. In the latter, agents employ a deep neural factorisation network to learn to assess their impact on global rewards. This approach delivers comparable performance for four agents and more, with a 34-fold speed improvement for 30 agents and only first-order computational time growth.

Finally, the impacts of implementing implicit coordination using these multi-agent reinforcement learning methodologies are modelled. It is observed that even without explicit grid constraint management, cooperating energy users reduce the likelihood of voltage deviations. The cooperative management of voltage constraints could be further promoted by the MARL policies, whereby their likelihood could be reduced by 43.08% relative to an uncoordinated baseline, albeit with trade-offs in other costs. However, while this thesis demonstrates the technical feasibility of MARL-based cooperation, further market mechanisms are required to reward all participants for their cooperation.

Multi-agent reinforcement learning for the coordination of residential energy flexibility

Flora Valérie Jeanne Charbonnier

Pembroke College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2023

Acknowledgements

I am lucky to have received support from numerous individuals on this journey. I would like to express my thanks to some of them below.

First, I would like to extend my sincere gratitude to my supervisors. Dr Thomas Morstyn's research first sparked my interest in distributed energy resources coordination, and our initial conversations were decisive in developing and instilling confidence in this project. Thomas is thorough, diligent, and meticulous in his approach. His dedication to ensuring the highest standards of work has been invaluable throughout the completion of this thesis. Prof Malcolm McCulloch has established an exceptionally supportive, collaborative and inclusive environment at the Energy and Power group, and he is genuinely dedicated to the development of his students and team members. He possesses a remarkable capacity for strategic insight, adeptly shaping the overarching vision of research. I count myself lucky to have benefited from his breadth of vision and depth of experience.

I am most indebted to the European Saven scholarship, established by Mr Bjorn Saven, for funding my research.

I am also grateful to my colleagues at EPG for their insights and support. It was truly special to grow with them as a researcher, scholar and individual. I want to thank Dr Miriam Miriam Zachau Walker and Dr Liyang Han for paving the way and providing wise advice as the veterans of the DPhils as I first joined; to Dr Alycia Leonard, Dr Aniq Ahsan and Claire Halloran for our rich conversations and collaborations, both in friendship and academia; to Dr Katherine Collett and Dr Stephanie Hirmer, for being leaders to look up to; to Dr Nicole Miranda, Dr Scot Wheeler and Dr Filiberto Fele for always making time to share their thoughts and provide help to their colleagues, with boundless patience and generosity; to Farhad Billimoria who shared my DPhil journey from down under. I am incredibly fortunate to have had the opportunity to learn from such brilliant individuals.

I was fortunate to have a network of exceptionally brilliant collaborators for my thesis research, beyond my supervisors. Working alongside Julie Vienne, who conducted her Master's thesis under my supervision in Oxford, was a thoroughly enjoyable experience. Julie has a strong appetite for learning, remarkable analytical capabilities, and a collaborative spirit. I am also thankful to have collaborated with her co-supervisors, Dr Eleni Stai and Ognjen Stanojev. Gaining insights from

such knowledgeable and committed individuals has been truly enriching. I also had the privilege of collaborating with Dr Bei Peng, who is a fountain of knowledge on multi-agent reinforcement learning, and a steadfast partner in our research pursuits, demonstrating exceptional insight, discernment, and precision.

I would like to thank Dan Cerigo for believing in and supporting me in my journey as a data scientist.

My appreciation extends to my friends, old and new, whose support does not go unnoticed. While engineering is my academic anchor, building links beyond its borders across the University of Oxford was particularly enriching. I particularly cherished my community at Pembroke College. When delving deeply into one's specific area of expertise and getting immersed in lines of code, it is rejuvenating to remember that one's research is just a small part of a much larger puzzle.

I thank my parents for nurturing my intellectual curiosity and instilling in me the self-confidence to pursue my studies and for their unwavering support throughout. They have patiently and lovingly provided encouragement, especially as I faced personal and health hurdles, until I could get back on my feet and complete this thesis. I am grateful for my sisters Victoire and Marie, who both keep me rooted and inspire me to reach out and spread my branches far and wide. Thank you to Annette, David and Christian for their warmth and constant encouragement. They have been my anchor and my home in the UK.

Finally, I must thank my wonderful husband, Luke. He provided love, support, and an endless stream of delicious homemade meals to sustain me through writing this dissertation. His immense knowledge is always an inspiration to build bridges beyond the boundaries of my research. I am truly fortunate to have him by my side.

Abstract

This thesis investigates whether residential energy flexibility can be coordinated without sharing personal data at scale to achieve a positive impact on energy users and the grid.

To tackle climate change, energy uses are being electrified at pace, just as electricity is increasingly provided by non-dispatchable renewable energy sources. These shifts increase the requirements for demand-side flexibility. Despite the potential of residential energy to provide such flexibility, it has remained largely untapped due to cost, social acceptance, and technical barriers. This thesis investigates the use of multi-agent reinforcement learning to overcome these challenges.

This thesis presents a novel testing environment, which models electric vehicles, space heating, and flexible household loads in a distribution network. Additionally, a generative adversarial network-based data generator is developed to obtain realistic training and testing data. Experiments conducted in this environment showed that standard independent learners fail to coordinate in the partially observable stochastic environment. To address this, additional coordination mechanisms are proposed for agents to practise coordination in a centralised simulated rehearsal, ahead of fully decentralised implementation.

Two such coordination mechanisms are proposed: optimisation-informed independent learning, and a centralised but factored critic network. In the former, agents lean from *omniscient* convex optimisation results ahead of fully decentralised coordination. This enables cooperation at scale where standard independent learners under partial observability could not be coordinated. In the latter, agents employ a deep neural factorisation network to learn to assess their impact on global rewards. This approach delivers comparable performance for four agents and more, with a 34-fold speed improvement for 30 agents and only first-order computational time growth.

Finally, the impacts of implementing implicit coordination using these multi-agent reinforcement learning methodologies are modelled. It is observed that even without explicit grid constraint management, cooperating energy users reduce the likelihood of voltage deviations. The cooperative management of voltage constraints could be further promoted by the MARL policies, whereby their likelihood could be reduced by 43.08% relative to an uncoordinated baseline, albeit with trade-offs in other costs. However, while this thesis demonstrates the technical feasibility of MARL-based cooperation, further market mechanisms are required to reward all participants for their cooperation.

Contents

List of Abbreviations

General abbreviations

ADMM . . . . Alternating Direction of Multipliers

AI . . . . . Artificial Intelligence

ANN . . . . . Artificial Neural Network

API . . . . . Application Programming Interface

CLNR . . . . . Customer-Led Network Revolution Dataset

CTDE . . . . . Centralised Training with Decentralised Execution

Dec-POMDP Decentralised Partially Observable Markov Decision Process

CVaR . . . . . Conditional Value at Risk

DER . . . . . Distributed Energy Resource

DLMP . . . . . Distribution Locational Marginal Pricing

DLT . . . . . Distributed Ledger Technology

DQN . . . . . Deep Q Networks

DRL . . . . . Deep Reinforcement Learning

DSO . . . . . Distribution System Operator

DSR . . . . . Demand-Side Response

EMS . . . . . Energy Management System

EV . . . . . Electric Vehicle

FACMAC . . . Factored Multi-Agent Centralised Policy Gradients

FLOP . . . . . Floating-Point Operation

GAN . . . . . Generative Adversarial Network

GNN . . . . . Graph Neural Network

GDPR . . . . . General Data Protection Regulation

HEDGE . . . . Home Energy Data Generator

HVAC . . . . . Heating, Ventilation, and Air Conditioning

IA2C . . . . . Independent Synchronous Advantage Actor-Critic

IEEE . . . . . Institute of Electrical and Electronics Engineers

ICE . . . . . Internal Combustion Engine

ICT . . . . . Information and Control Technology

IPPO . . . . . Independent Proximal Policy Optimization

IQL . . . . . Independent Q-Learning

KKT . . . . . Karush–Kuhn–Tucker

LQR . . . . . Linear Quadratic Regulator

MAA2C . . . . Multi-Agent Advantage Actor-Critic

MADDPG . . Multi-Agent Deep Deterministic Policy Gradients

MAMDP . . . Multi-Agent Markov Decision Process

MAPPO . . . Multi-Agent Proximal Policy Optimization

MARL . . . . Multi-Agent Reinforcement Learning

MDP . . . . . Markov Decision Process

MERRA . . . Modern-Era Retrospective Analysis for Research and Applications

ML . . . . . Machine Learning

MPC . . . . . Model Predictive Control

NASA . . . . . National Aeronautics and Space Administration

NTS . . . . . National Travel Survey

OPF . . . . . Optimal Power Flow

POMDP . . . Partially Observable Markov Decision Process

p.u. . . . . . per unit

PV . . . . . Photovoltaic

P2P . . . . . Peer-to-Peer

RL . . . . . Reinforcement Learning

ROPF . . . . . Relaxed Optimal Power Flow

SDG . . . . . Sustainable Development Goal

TOU . . . . . Time of Use

TRTS . . . . . Train on Real, Test on Synthetic

TSO . . . . . Transmission System Operator

TSTR . . . . . Train on Synthetic, Test on Real

UK . . . . . United Kingdom

VDN . . . . . Value-Decomposition Networks

V2G . . . . . Vehicle-to-Grid

. . . . . .

Independent Q-Learning Methodologies

AE . . . . . Use Advantage rewards, learn from Environment

AO . . . . . Use Advantage rewards, learn from Optimisations

ME . . . . . Use Marginal rewards, learn from Environment

MO . . . . . Use Marginal rewards, learn from Optimisations

TE . . . . . Use Total rewards, learn from Environment

TO . . . . . Use Total rewards, learn from Optimisation

. . . . . .

Reliability Metrics

DT . . . . . Dispersion across Time

DR . . . . . Dispersion across Runs

LRT . . . . . Long-term Risk across Time

RR . . . . . Risk across Runs

SRT . . . . . Short-term Risk across Time

List of Symbols

Roman symbols

a_i^t	RL action
$a_{\mathbf{EV},i}^t$	RL action controlling the EV battery
$a_{\mathbf{heat},i}^t$	RL action controlling the heating
$a_{\mathbf{cons},i}^t$	RL action controlling the household electric consumption
$a_{\mathbf{reactive},i}^t$	RL action controlling the EV battery reactive power injections
$b_{\mathbf{in},i}^t$	Charge into the battery [kWh]
$b_{\mathbf{out},i}^t$	Discharge out of the battery [kWh]
$c_{\mathbf{b}}^t$	Battery costs [£]
$c_{\mathbf{d}}^t$	Distribution costs [£]
$c_{\mathbf{g}}^t$	Grid costs [£]
$c_{\mathbf{v}}^t$	Voltage costs [£]
$\hat{c}_{i,k,t_{\mathbf{C}},t_{\mathbf{D}}}$	Household electric consumption at time t_{C} by prosumer i for load of type k (fixed or flexible) demanded at t_{D} [kWh]
$\mathbf{C}_{\mathbf{b}}^t$	Battery depreciation coefficient [£/kWh]
$\mathbf{C}_{\mathbf{d}}^t$	Distribution charge on exports [£/kWh]
$\mathbf{C}_{\mathbf{g}}^t$	Grid cost coefficient [£/kWh]
$\mathbf{C}_{\underline{\mathbf{v}}}$	Under-voltage cost coefficient [£/kWh]
$\mathbf{C}_{\overline{\mathbf{v}}}$	Over-voltage cost coefficient [£/kWh]
$\mathbf{d}_i^t$	Household electric demand [kWh]
$\mathbf{d}_{\mathbf{EV},i}^t$	EV demand for required trips [kWh]
$\mathcal{D}$	GAN discriminator model
E_i^t	Battery energy levels [kWh]
f_t	HEDGE profile scaling factor
F	Objective function

$\hat{F}_t$	Part of the system objective function corresponding to time t
g^t	Total grid import [kWh]
g_ψ	Non-linear monotonic function parameterised as a mixing network with parameters ψ
$\mathcal{G}$	GAN generative model
G	Discounted return
h_i^t	Electric heating consumption [kWh]
I_{jk}	Squared current [A^2]
$\mathbf{K_{fixed}}$	Matrix of connection of fixed buses on the network
$\mathbf{K_{flex}}$	Matrix of connection of flexible buses on the network
l_i^t	Household loads [kWh]
M	Large number
o_i	Agent observation
$O(n), O(n^2)$	Big O notation for algorithm run time (first order, second order).
p_j^t	Bus-level real power import [kWh]
p_D	Discriminator network dropout rate
p_G	Generator network dropout rate
$\mathbf{p}^t_{\mathbf{PV},i}$	PV production [kWh]
P_{jk}	Real power flow [kWh]
P	State transition function
P_κ	Matrix of behaviour cluster transition probabilities
P_f	Matrix of scaling factor transition probabilities
PF	Power factor
q_j^t	Bus-level reactive power import [kVAR]
$q^t_{\mathbf{EV},i}$	Electric vehicle battery reactive power import [kVAR]
$Q(s,a)$	Q-table (state-action values)
r^t	RL reward [£]
$\mathbf{R}$	Average resistance between the main grid and the distribution network [Ω]
s_i^t	RL state $s_i^t \in \mathcal{S}$
$S_{\mathbf{EV}}$	Maximum apparent power capacity of the EV battery [VA]

$\underline{\mathbf{T}}_i^t$ . . . . . Lower bound on indoor air temperature [^{o}C]

$\overline{\mathbf{T}}_i^t$ . . . . . Upper bound on indoor air temperature [^{o}C]

$T_{\mathbf{air},i}^t$ . . . . . Indoor air temperature [^{o}C]

$\mathbf{T}_{\mathbf{e}}^t$ . . . . . External temperature [^{o}C]

$T_{\mathbf{m},i}^t$ . . . . . Building mass temperature [^{o}C]

$\underline{\mathbf{v}}$ . . . . . Minimum voltage level [V]

$\overline{\mathbf{v}}$ . . . . . Maximum voltage level [V]

v_j . . . . . Voltage levels [V]

v_j . . . . . Voltage levels [V]

$V(s)$ . . . . . State-value function

w_j^t . . . . . Squared voltage [V^2]

W . . . . . Loss weight

$\mathbf{x}_{jk}$ . . . . . Line reactance [Ω]

. . . . . .

Indexes . . . .

i . . . . . Index of homes $i \in \mathcal{N}$

j . . . . . Index of buses $i \in \mathcal{B}$

j, k . . . . . Index of pairs buses $(j, k) \in \mathcal{L}$

m . . . . . Load type (fixed or flexible)

t . . . . . Index of time steps $t \in \mathcal{T}$

u . . . . . Index of profile

$t_{\mathbf{C}}$ . . . . . Consumption time step

$t_{\mathbf{D}}$ . . . . . Demand time step

. . . . . .

Sets . . . . . .

$\mathcal{A}$ . . . . . Set of actions

$\mathcal{B}$ . . . . . Set of buses

$\mathcal{L}$ . . . . . Set of lines

$\mathcal{N}$ . . . . . Set of homes

$\mathcal{O}$ . . . . . Set of agent observations

$\mathcal{R}$ . . . . . Real numbers

$\mathcal{S}$ . . . . . Set of states

$\mathcal{T}$ . . . . . Set of time steps

. . . . . .

Greek letters .

α . . . . . Learning rate

β . . . . . Hysteretic rate

δ . . . . . Temporal difference error

ℓ . . . . . Loss

ϵ . . . . . Exploration

$\epsilon^t_{\mathbf{ch},i}$ . . . . . Battery charging losses [kWh]

$\epsilon^t_{\mathbf{dis},i}$ . . . . . Battery discharging losses [kWh]

$\epsilon^t_{\mathbf{g}}$ . . . . . Main grid losses [kWh]

$\eta_{\mathbf{ch}}$ . . . . . Charging efficiency

$\eta_{\mathbf{dis}}$ . . . . . Discharging efficiency

γ . . . . . Depreciation rate

κ . . . . . Behaviour cluster

$\lambda_{i,k,t_{\mathbf{C}},t_{\mathbf{D}}}$ . . . . Flexibility boolean matrix

μ^t_i . . . . . EV at-home availability boolean

ν . . . . . Day type (weekday or weekend day)

ξ . . . . . Recursive linear heating model coefficients matrix

π . . . . . RL policy

τ_i . . . . . Local action-observation history

ρ_{jk} . . . . . Line resistances [Ω]

Φ^t . . . . . Solar heat flow rate [W]

$\boldsymbol{\phi}$ . . . . . Parameters of the centralised action-value critic function Q^{π}_{tot}

ϕ_i . . . . . Parameters of the agent-wise utilities $Q^{\pi_i}_i$

ψ . . . . . Parameters of mixing network

σ . . . . . Standard deviation

θ . . . . . Parameters of actor

Θ . . . . . Replay buffer

Ω_{jk} . . . . . Reactive power flow [kVAR]

1 Introduction

Contents

This thesis considers the research question: Can residential energy flexibility be coordinated without sharing private data at scale to achieve a positive impact on energy users in the electricity grid? For this purpose, a fully decentralised cooperation approach is pursued using multi-agent reinforcement learning (MARL).

1.1 Context and motivation

This section first establishes the background and rationale for this research.

1.1.1 The rising need for demand-side response

To help tackle climate change, greenhouse gas emissions must be reduced drastically and urgently. This is crucial to limit anthropogenic global warming to 1.5°C above pre-industrial levels and reduce risks of climate-related impacts on health, livelihood, security and economic growth. Particularly, energy systems represent three quarters of global greenhouse gas emissions [1]. Therefore, their decarbonisation is a priority.

The current technological options for decarbonised energy supply include renewable and nuclear energy and use electricity as a vector. However, as of 2022, only one fifth of final energy consumption in the world uses electricity [2]. As a result, we must follow a two-pronged strategy for the decarbonisation of energy systems: widespread electrification of primary energy provision and decarbonisation of the power sector [3]. Beyond mitigating climate change, this electrification of energy usages and decarbonisation of electricity supply will have numerous health, environmental[1] and economic[2] benefits.

However, the widespread electrification of energy use, concurrent with the increase in renewable energy penetration, could cause commensurable disruption without appropriate management. Not only is the total electricity consumption expected to more than double between 2020 and 2050 [6], but intermittent and distributed renewable power supplies could be required to supply 70% to 85% of electricity by 2050 [7]. These shifts pose the challenges of the intermittency and limited controllability of resources [8].

As opposed to water and gas, which can be stored in network pipelines to some extent, electricity networks[3] do not offer such buffering. Grid balancing is therefore needed to ensure that electricity consumption matches electricity production at any moment. Otherwise, frequency levels deviate from their specified levels, causing generator outages or blackouts. Historically, this balance has been maintained by

[1] In some parts of Europe, household heating with solid fuels (including coal and biomass) contributes to as much as 75% of outdoor fine particulate matter pollution [4]

[2] Nearly two-thirds of newly installed renewable power in 2021 had lower costs than the world's cheapest coal-fired options in the G20 [5]

[3] "Network" is henceforth used to refer to the electricity network interchangeably to improve readability

fossil-fuel-based *dispatchable* generation technologies such as coal and gas power plants, which can be programmed on demand according to system needs. As electricity generation is increasingly non-dispatchable, a paradigm shift towards demand-side response (DSR) is needed to match our consumption to the available generation and maintain grid stability.

The next challenge for energy systems will thus not be that of energy source availability, as renewable resources are plentiful, but that of low-carbon flexibility provision. Coordinating flexible distributed energy resources (DERs) such as those described in Section 1.1.2 can help reduce the costs of centralised transmission, storage, peaking plants and capacity reserves, improve grid stability, align demand with decarbonised energy provision, promote energy independence and security, and lower household energy bills [9–11].

1.1.2 The flexibility potential of residential energy

Residential sites constitute a significant share of potential DSR, representing, for example, 38.5% of the 2019 UK electricity demand and 56.4% of energy consumption if including transport and heat, which are both undergoing electrification [12]. The Committee on Climate Change estimates that as much as 53% of household demand could be flexible in the future [13]. DSR potential is broadly proportional to the peak electricity demand [14], of which British households already represent about 50% [15].

The residential energy flexibility potential comprises various DERs. This thesis focuses on the three flexible technologies listed below, which are already expected to be available in buildings by 2050, independently of additional investments for flexibility purposes:

1. Electric vehicles (EVs): A shift is occurring from oil to electricity as an energy source for private transport. Increasing ownership of EVs has been facilitated by plummeting costs, with an 82% levelised cost drop between 2010 and 2022[4] [16, 17], and promising prospects of further declines in the coming years. Regulation changes further facilitate the switch to electric cars, with a

[4]EV batteries went from over 1,100 $/kWh in 2010 to 151$/kWh in 2022

growing list of countries pledging to phase out internal combustion vehicles, for example banning their sale from 2035 in the UK. If half of British cars switched to EVs, the instant power capacity added would be three times the UK's 2022 peak electricity demand. This capacity would be equivalent to one average day of electricity demand[5], though with intermittent availability as cars are not always connected to the network. While EVs have high power consumption, they also have charging time flexibility thanks to their storage capacity, which can temporally decouple generation and consumption. Given that the investment in these batteries will already occur for transport purposes, their use should be optimised for flexibility provision, with benefits for both asset owners and the network.

2. Flexible household loads: A large share of household loads can be flexible. While this thesis considers that cooking, lighting, and electronics are not flexible as displacing their use would interfere with the livelihood of residential users, wet (e.g. washing machine, dryer, dishwasher) and cold (e.g. freezer, refrigerator) appliances represent 18 and 15% of UK domestic electricity consumption in 2012 respectively, excluding space and water heating. Considering that these can often be deferred without impacting home inhabitants, a third of household loads are estimated to be partially flexible. This share is expected to remain constant, with wet and cold loads to represent 25 and 11% of the UK annual domestic electricity demand in 2030 [21]. The advancement of power electronics, communication, and control technologies has facilitated the enhancement of electricity consumption dynamicity. This development allows flexible demand to be used as DERs in many contexts [22].

3. Heating: As a shift is operated from gas boilers to electric heat pumps, the potential of thermal storage as a source of flexibility is sizeable in cold and temperate climates. For example, in the EU, cooling and heating already

[5] considering the 39.5 million car fleet in the UK [18], electric cars with 7 kW charging rate and 50 kWh storage capacity not used for mobility, the winder 2022/2023 peak electricity demand of 45.3 GW [19] and an average day of electricity demand 0.95 TWh [20]

represent nearly half of the energy consumption[6], and smart residential storage heaters and hot water cylinders represent nearly four times the dedicated storage capacity such as that offered by pumped hydro. By 2050, up to 51GW of controllable demand is projected across the EU [23], especially as heating and cooling efficiency, affordability and sustainability are key policy objectives. Adopting heat pumps could however double electricity peak demand in the UK [24], especially as heat pump loads are correlated with the weather and, consequently, highly correlated together. Given the increasing interdependencies between the heat and electricity energy systems [25], these loads will add pressure on the electricity grid, whilst simultaneously offering opportunities to provide flexibility for the electricity system [4].

Moreover, photovoltaic (PV) panels have become increasingly affordable, with an 88% levelised cost drop between 2010 and 2021 [5], opening the way for the distributed ownership of the technology. Though this energy source is not dispatchable and will not be considered flexible in this thesis, it has some flexibility in its potential curtailment (downward flexibility). Moreover, modelling its adoption is critical not only as a flexible resource but also as a resource requiring associated flexibility to use the decarbonised energy it supplies efficiently.

So far, the coordination of the flexibility offered by these resources is very limited. This lack of coordination may cause issues for the network, as heat and transport are being electrified and distributed PV panels adopted. Uncoordinated EV charging could for example cause the peak electricity demand to increase by 35% in the UK [26] with adverse consequences on the electricity system, in particular on the low-voltage distribution networks [27, 28]. The network reinforcement costs could be considerable, even more so if DERs are uncoordinated [29]. DNOs may thus deny permissions for installations of, or even demand the disconnection of, uncoordinated DERs. Therefore, even without mature market mechanisms for the remuneration of residential energy flexibility provision, it may be in the DER's owner's interest to manage their impacts on the system cooperatively.

[6] 50% (546 Mtoe) of final energy consumption in 2012

Using demand-side coordination of these locally available sources of flexibility can lower the costs of decarbonisation by minimising the disruption caused by adopting variable energy sources and aligning demand with decarbonised energy provision. Employing the flexibility readily available in residential homes can reduce the need to invest in and operate other types of costly flexibility, such as standalone lithium-ion batteries and pumped-storage hydroelectric dams. Moreover, DSR can help reduce peak demand – up to 75% of the increase in peak electricity demand due to EV charging could be offset using smart charging alone [26]. This is crucial as, generally, 20% of the power generation capacity is latently available to meet the peak demand that occurs for approximately 5% of the time [30]. DSR can help meet new demands with variable energy sources without adding new generation and network capacity, limiting the need for network reinforcement, peaking plants and capacity reserves.

Moreover, in a time of energy affordability crisis[7], the coordination of residential energy flexibility can help lower household energy bills in two ways. Firstly, arbitrage opportunities arise when exposed to variable tariffs. Secondly, reducing overall systems costs can indirectly benefit customers, who eventually bear these costs. Other system benefits include improving grid stability, promoting energy independence and security, reducing electricity network losses, and increasing its resilience [4, 9, 10, 29]. As a result, the involvement of energy users in providing flexibility to allow for the integration of variable and distributed renewable electricity generation is a key policy objective to meet climate targets. This is illustrated by the efforts underway to create market frameworks to reward flexibility provision [32, 33].

1.2 The challenges and requirements of residential energy coordination

The motivation for the use of residential energy flexibility was explained in Section 1.1. This section now delves into the challenges of its coordination. These

[7]The post-pandemic economic rebound has outpaced energy supply, which has escalated into a widespread global energy crisis following the 2022 Russian invasion of Ukraine. As of December 2022, it was estimated that some 6.7 million UK households are for example in fuel poverty [31]

challenges to overcome will correspondingly lead to the thesis's research questions and coordination success criteria in Section 1.3.

The flexibility potential of residential energy described in the previous section is so far under-exploited, as implemented DSR programmes primarily focus on larger, well-known industrial and commercial actors that require less coordination and data management [34], with most residential customers still limited to trade with utility companies [35]. As laid out below, the primary hurdles to unlocking residential flexibility are the cost of system transition as the domestic potential is highly fragmented [14], concerns about privacy and hindrance of personal activities [10, 36], computational challenges for real-time control at scale [37], and the need to mitigate the risks of potential negative impacts.

1.2.1 The cost of system transition

The cost of system transition has been a hindrance to residential DER coordination. This includes both infrastructure installation costs and market transformation hurdles, which limit the enabling environment for residential energy coordination.

Firstly, the high costs of two-way communication and control equipment installation have represented a major barrier to small energy user DSR [38]. These costs are increased by the high fragmentation of the residential energy flexibility potential. There were, for example, an estimated 27.8 million households in the UK in 2020 [39], each home with relatively low flexibility potential [14]. Moreover, residential users present substantial heterogeneity, and the roll-out of advanced information and control infrastructure (ICT) only yields local economic potential in a fraction of cases [14]. Not every household may be able to justify the investment and maintenance costs of such infrastructure in [40–42], especially in the context of the current energy crisis. In addition, even for households that may present financial benefits, liquidity[8] for upfront costs may be prohibitively high for non-professional actors [43].

[8] access to readily available capital or financing options

In addition, existing market and regulatory frameworks are ingrained. Numerous commercial peer-to-peer projects[9] have been interrupted, not only due to infrastructure and logistical costs, but also the risks of high regulatory uncertainty, market barriers and specialist resource commitment [44, 46]. Creating new market structures incurs high upfront costs to develop new tariffs, manage debt pathways, customer service, support and onboarding, IT systems costs, data compliance costs, communication between different suppliers, and risks of additional complexity creating consumer confusion and misuse [45]. Some experts thus believe an overhaul of the current centralised energy grid is unlikely to happen within the next decade [47].

Therefore, this research will develop a simulation framework to assess whether value can be obtained within existing structures. It will experiment with flexibility assets readily available in the residential sector, with no or minimal further investment. Namely, it will model the buildings' thermal mass and electric vehicles, as investment in these assets will occur regardless of whether they are used for flexibility management. Furthermore, this thesis will model homes without access to real-time communication infrastructure and networks, leading to a partial information coordination problem. Finally, in the short and medium terms, innovative solutions must fit with the existing highly regulated energy markets and rely on the existing infrastructure and incumbent energy players [44].

1.2.2 Acceptability

While there is a strong motivation for taking personal action to help fight climate change [48, 49], people engaging in demand-side response in a domestic setting may have concerns about privacy, comfort, and agency [10, 36].

Privacy protection results in a trade-off between control optimality and data availability. On the one hand, lower degrees of data availability increase a control strategy's optimality gap due to the missing information problem. On the other hand, as sensitive information about users' habits and lifestyles can be inferred from

[9] e.g. such as those led by Piclo, Pylon Network, Verv, Repowering, Elexon [44] and the ESB Networks Dingle P2P trial [45]

electricity data [50, 51], perfect knowledge of a user's power consumption profile by a utility may be considered a violation of privacy. This privacy breach may damage the trust required to implement demand response systems [52]. Moreover, protecting privacy is not solely a matter of trust and ethics but also helps manage regulatory risks. After decades of regulatory lag around individuals' privacy and consent-based data sharing [53], regulators are starting to catch up [54], with increasing momentum towards stronger regulations to protect individuals' right to data ownership [55–57].

Comfort is a primary consideration for potential DSR participants [42] – a field evaluation concluded that thermal comfort is more important to people than energy efficiency, even for affordable housing residents who pay their own heating bills[10] [58]. Energy consumption is not a social practice in itself but is rather an enabler of practices such as cooking, showering or driving [10]. Interference must be limited, as changes in consumption patterns, interference with temperature set-points, and required efforts to acquire information all cause dissatisfaction [9]. Affecting occupants' comfort makes them less likely to engage with demand response schemes [15]. This aversion to external interference is also linked to an inclination for a sense of agency in one's living space and a reluctance to relinquish control of appliances in one's own home to an external entity [36, 59]. As argued by Darby et al. [15], "the ability to control can itself contribute to a sense of well-being and comfort".

However, potential adopters paradoxically favour minimal involvement over active control of their energy management [10]. Participants may have gaps in the knowledge required to participate most effectively in load-shifting programmes or make unsustainable efforts to provide limited flexibility, resulting in wasted efforts and inefficient energy consumption patterns[11]. Relatively few consumers would merely change suppliers to obtain more favourable tariffs, even when it may appear to be a straightforward choice [61]. This hesitancy suggests there may be a lack of interest or understanding regarding more complex energy management strategies

[10] Though energy prices have increased since 2012, when this field evaluation was conducted, so this conclusion may not hold today.

[11] By way of illustration, participants in a trial were reported to switch their mobile chargers off – which contributes about three orders of magnitude less load shifting than white goods and at higher disturbance – or to stay up at night to shift tumble dryer use [60].

[14]. Moreover, individuals encountering a sense of disorientation in relation to the technological infrastructure in their household may experience a diminished sense of control, which can be a source of distress. The literature is thus consistent that simple electricity pricing programs are preferred by energy users [60, 62–64]. This raises the question of the feasibility of the active participation of energy users, especially if sophisticated real-time optimisation of their appliances is required.

This thesis, therefore, asks whether worthwhile coordination can be achieved that does not impede users' privacy, comfort and agency. Coordination strategies will be developed that keep all personal information and control within one's home. Agency will be maintained as the household activities, and trips and heating requirements taken as fixed inputs to the control model. Only residential loads whose control can be automated without impacting participants' comfort or daily activities will be deemed flexible – which also happen to be those with the highest flexibility potential. As such, the battery of EVs can be used between trips so long as they are suitably charged at the time of trips. Thermal loads are suitable for demand response thanks to the inertia of the heat in the home, which allows for short-term interruptions while maintaining thermal comfort according to the users' instructions. Finally, there is a need for a third way beyond either external control or manual energy management, where DSR is fully automated without requiring excessive individual involvement, while keeping the control decentralised within individual homes.

1.2.3 Computational scale

Computational scale is a hurdle for the direct control of the millions of residential DERs in a country.

One traditionally used mathematical framework for computing optimal power dispatch [65] is convex optimisation. It guarantees the maximisation of global coordination objectives in convex problems with variables known ahead of time [66]. However, residential energy coordination presents challenges to its application due to the ICT and privacy issues mentioned above, which result in a lack of centralised information, and to the limited scalability of centralised computations. These

challenges will be laid out in Section 2.2.1. One could consider an alternative based on bilateral communication. However, as discussed in Section 2.1.3.2, an iterative approach based on dual price variable adjustment can pose challenges, such as the number of bilateral iterations until convergence. Computational issues may arise as the complexity increases with the number of DERs at scale [67].

Alternative coordination computation frameworks must consequently be developed. There is a need for coordination algorithms that limit the use of communication networks, distributed computing, and cloud computing at scale, as highlighted below.

1. The costs of communication networks and distributed computing are significant, with a need for sufficient and predictable network latency and bandwidth if communicating large amounts of residential energy data, as well as a need to manage potential related security and privacy risks [68]. Previous domestic demand response trials have further identified how connectivity, i.e. communication between technologies, including existing systems, was a major hurdle for the successful completion of projects, with significant human and technical resources spent on connectivity between individual hardware and software components [15]. Integration into highly complex systems may provide control benefits but increases the risk of failure [42].

2. The use of extensive decentralised computational resources is not feasible in the context of residential energy as, as of 2023, individual homes have limited computation resources. Moreover, individual homes may not have the liquidity or sufficient incentives to invest in powerful computation resources. Computations during implementation should thus be light enough to be computed with low-cost equipment at the home level.

3. While burdensome computations could be performed in the cloud, the use of computational resources has material impacts on the depletion of finite resources – and associated human conflicts – on soils, rivers, landscapes, ecological systems, atmosphere and the climate [53]. Coordination algorithms

that aim to contribute towards decarbonisation efforts must therefore limit their computational burden at scale.

Therefore, this thesis will seek to make the training of reinforcement learning policies tractable and their implementation feasible with limited local computational resources.

1.2.4 Risks of negative impacts

This thesis considers an engineering problem that could extend beyond theory to reach human subjects and physical infrastructure in the near future. As artificial intelligence reaches all corners of human societies "[the] separation of ethical questions away from the technical reflects a wider problem in the field, where the responsibility for harm is either not recognised or seen as beyond the scope of the research" [53]. The implementation of AI algorithms into homes and electricity networks requires testing to ensure it does not increase the likelihood of harm to energy users or the grid.

In the context of electricity network applications, the presence of a large number of DERs can pose challenges to the stability and reliability of the critical distribution network infrastructure [69]. This can lead to over-voltages, under-voltages, congestion [70] and phase imbalances, which negatively affect power quality [71]. However, if properly managed, DERs have the potential to offer valuable services to Transmission System Operators (TSOs) and Distribution System Operators (DSOs) [72].

As for energy users, models of energy collectives show that there is the potential to gain from cooperating for both the community as a whole and each individual prosumer[12], but only they are designed adequately [37]. The implementation of P2P markets can result in the emergence of energy poverty among consumers or communities with limited economic resources [33]. This is particularly true when taking a statistical approach such as RL to maximise expected outcomes, with no

[12]Prosumers are proactive consumers with distributed energy resources actively managing their consumption, production and storage of energy [73]. In this thesis, "home", "agent", "energy user" and "prosumer" are used interchangeably.

hard theoretical guarantees on outcomes. Risk-neutral control RL fails to account for the probability of deviating significantly from the expected outcomes, whether in a negative or positive direction [74]. Due to the inherent stochastic nature of the problem, even a policy that is considered optimal may exhibit subpar performance in some instances [75]. Moreover, maximising total system utility may come at the expense of individual energy users [76]. This would be unacceptable both in terms of meeting the sustainable development goal (SDG 7) of access to affordable, reliable, sustainable and modern energy for all [77] and in terms of fairness [37]. Energy regulators advise that users should contribute fairly towards network costs and receive compensation that reflects their contribution towards them [78].

Therefore, this thesis will seek to assess whether the algorithms achieve positive impacts for energy users and the grid, or whether there are risks of distribution network voltage deviations and of increased costs for energy users resulting from the coordination of residential energy flexibility.

1.3 Approach to addressing the research question

The coordination of residential energy flexibility was motivated in Section 1.1, and the specific hurdles to its realisation were laid out in Section 1.2. Correspondingly, this section presents the core research question investigated in this thesis, along with its subsidiary research questions. Furthermore, each question is accompanied by a clear success criterion, which will serve as a measure to evaluate the research outcomes.

Main research question: *Can residential energy flexibility be coordinated without sharing private data at scale to achieve a positive impact?*

This overarching research question is broken down into four subsidiary research questions:

1. Can the efficacy of algorithms that coordinate residential energy flexibility be assessed?

Success criterion: In order to answer this question, a testing framework specific to residential DSR coordination algorithms must be developed. To meet the specific constraints set out in Section 1.2.1, it must model the local-level control of existing DERs (intermittently available EVs, space heating, household holds, PV) in response to existing energy tariffs in the form of one-way communication signals to the users. The experimenting and testing environment must include home and network modelling as well as the generation of sufficient realistic input data for robust training and testing.

2. Can residential energy flexibility be successfully coordinated without sharing private data?

   Success criterion: As set out in Section 1.2.2, the need to preserve privacy and local agency limits the access to centralised information and control of DERs. The aim is to limit the control of appliances to the local level, with no communication of private data, thermal discomfort, or hindrance and delay of activities, without relying on accurate individual forecasts or real-time central computations. Despite this information and control gap, cooperative multi-agent coordination algorithms implemented in a decentralised way without sharing private data should yield value relative to an uncoordinated system. This coordination performance should be maintained as the system size increases, measured in savings in energy, network and carbon costs obtained per home and month.

3. Can this be achieved in a computationally scalable manner?

   Success criterion: A successful coordination will be considered to have overcome the scalability challenges presented in Section 1.2.3, if it only imposes minimal and constant distributed computation burden during implementation as the system size increases up to at least the feeder level ($\sim$100 homes), and with only first-order computational requirement growth for training $O(n)$.

4. Can the algorithms achieve positive impacts for energy users and the grid?

   Success criterion: The risks of implementing AI systems in the electricity grid presented in Section 1.2.4 should be evaluated. This cooperation should not increase the likelihood of voltage deviations in the distribution network, nor increase private costs for individual energy users.

1.4 Scope

This thesis is constrained in its scope in the following ways:

1. In the context of this thesis, *flexibility* is defined as potential changes in consumption and power injections into the electricity network to help match electricity consumption and generation. Table 1.1 presents a classification of the different approaches to providing flexibility. This thesis focuses on the shaded cells, i.e. on moving electricity consumption in time either by delaying loads, or using the battery of electric vehicles. It does not include infrastructure-heavy options such as hydroelectric dams, nuclear power generation, standalone grid-scale batteries, and transmission networks, among others. We also do not consider renewable energy curtailment in this thesis.

2. The modelled technologies are assumed to be readily available in homes in the medium term. While there is an early trend for investment in static batteries in the residential sector, their impact is negligible at the network scale at this stage. As such, standalone batteries are not explicitly considered but could easily be included in the proposed framework as EV batteries that are always connected to the grid.

3. The systems costs include energy prices, local battery degradation costs, grid losses, distribution network export charges, network voltage deviations and the cost of greenhouse gas emissions to society. They do not include power thermal line constraints and phase imbalance.

Table 1.1: Ways of providing flexibility, i.e. to change the occurrence of consumption and power injections into the electricity network. Shaded cells denote the sources of flexibility considered in this thesis.

	In time	In space
Change occurrence of consumption	e.g. delay washing machine use, industrial loads interruption	e.g. distribute computations to data centres in different geographical locations based on renewable energy availability
Change occurrence of generation	Decarbonised options include hydroelectric dams and nuclear generation, with both upward and downward flexibility; Other renewable energy sources only have downward flexibility through curtailment	Not pursued in current energy systems
Move energy	Energy storage (e.g. fixed batteries, vehicle-to-grid (V2G))	e.g. Transmission grid

4. In terms of the time scales considered, this thesis considers an intra-day operation problem, as opposed to short-term ancillary service provision (e.g. frequency response), or to a long-term energy systems planning problem.

5. The settlement of bills is beyond the scope of this thesis, which focuses on the operational problem. Thus, only one-way communication of price signals is needed for real-time feed-forward coordination, with no sharing of private data. However, some feedback of information on past energy consumption will be required at slower time scales so that users can settle their bills. Only aggregated numbers will be required, as baselining can be computed locally.

6. This thesis takes a statistical approach to maximising the expectancy of global utility and minimising the likelihood of negative impacts given partial observability and uncertainty. It does not concern itself with an exact and optimal control problem with perfect knowledge and control of network and residential components.

7. The problem investigated is that of optimal cooperation given existing time-of-use tariffs. The design of market incentive mechanisms or game theoretical settlement frameworks is not considered.

8. The primary focus of this thesis is the design of MARL-based cooperation algorithms, given a system model, as opposed to novel modelling techniques. Scalability is often prioritised over accuracy. As such, the battery model selected is linear, including the battery losses and degradation model; state-of-charge dependencies are not considered.

9. British input data is used where relevant, though the methodology developed is agnostic to the application area. Relevant input data need only be changed to be applicable to other contexts.

10. Cooling loads are not modelled due to their low prevalence in the British context, but air conditioners are heat pumps in reverse, and the analysis could easily be extended to include cooling needs.

1.5 Thesis structure

The remainder of this thesis is structured as follows.

Chapter 2 starts by analysing the landscape of the literature connected to the problem tackled by this thesis and identifying research gaps to be bridged. The four following body chapters then address the research questions listed above.

To assess the efficacy of algorithms that coordinate residential energy flexibility, a testing environment needs to be developed. This novel environment described in Chapter 3 includes a generative adversarial network (GAN)-based home energy data generator. Next, Chapter 4 highlights the need for and develops coordination mechanisms to achieve cooperation between independent learners in a stochastic environment with partial observability. Building on the insights learned in this chapter, Chapter 5 uses a deep multi-agent reinforcement learning approach with a centralised but factored critic for improved computational efficiency at scale. Having

developed these algorithms, Chapter 6 evaluates the impact of their implementation on the electricity network and energy users in an indicative system. To answer the primary research question of this thesis, Chapter 7 finally concludes that residential energy flexibility can be coordinated without sharing private data at scale to achieve a positive impact, with further research needed to reward participation.

2
Literature review

Contents

This chapter presents the landscape of the literature connected to the problem tackled by this thesis in the fields of distributed energy resources in Section 2.1 and multi-agent reinforcement learning in Section 2.2. The research gap this thesis

aims to bridge is then presented in Section 2.3.

2.1 The landscape of distributed energy resources coordination

This section investigates how strategies for the coordination of grid-edge energy resources – DERs connected at the distribution network level [79] – can be synthesised in a hierarchical classification according to their structural similarities and differences, referred to as a taxonomy [80].

As stated in Chapter 1, the coordination of grid-edge resources could make a significant contribution to power system decarbonisation, which is critical for keeping anthropogenic warming below 1.5°C above pre-industrial levels. While numerous energy generation and enabling storage and flexibility technologies exist and are emerging, a key challenge is their integration at an unprecedented scale. The coordination of energy flexibility in all sectors is needed to integrate high levels of intermittent renewable energy, from the transmission to the distribution network [7].

As the flexibility, control, and data ownership are increasingly decentralised, DERs can provide such decentralised DSR, as well as operational services such as frequency regulation, spinning reserves provision, voltage management, system balancing and network congestion management [81, 82]. Moreover, the coordination of flexible DERs can yield numerous local and global co-benefits: reduction of environmental concerns, reduced energy transport and storage costs, improved grid stability, alignment of peak demand with decarbonised energy provision in time and space, reduced costs of peaking plants and capacity reserves, deferral of transmission and distribution grid upgrades, energy independence and security, incentives for the contribution of numerous actors in investment, reduced bills for consumers, and enhanced social cohesion [9, 10, 83, 84]. There is therefore an extensive literature detailing research to extend the realm of coordination to small, grid-edge resources in the distribution grid, and the scholarship on the coordination of resources at the edge of the electricity grid has been growing exponentially since 1995 (Figure 2.1). Particularly, novel research is seeking to tackle the challenges of computation and

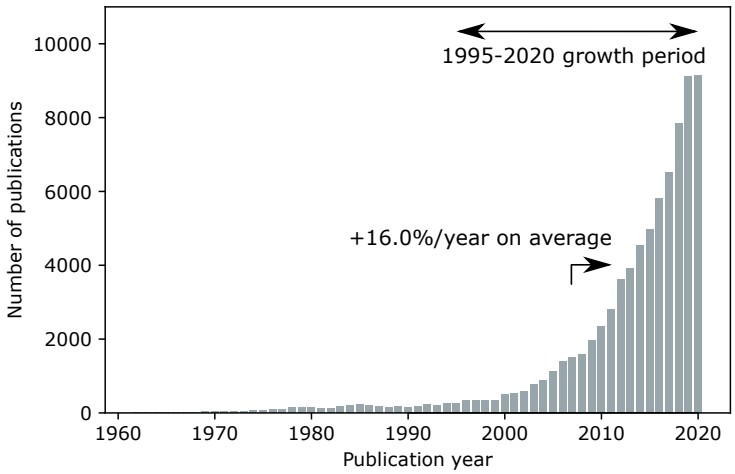


Figure 2.1: Number of publications in a selected body of literature on the coordination of electricity grid-edge resources between 1962 and 2020. See Appendix A.1 for the details of the systematic literature search. The number of publications grew exponentially by 16.0% per year on average between 1995 and 2020.

control at the scale of millions of units [37], privacy concerns and acceptability issues [9, 10, 42], as well as increased uncertainty at the local level [8].

Although numerous reviews focus on specific areas within the field of grid-edge energy resources coordination, they concentrate on limited aspects rather than systematically reviewing the landscape of coordination strategies, and use conflicting terminology. Some reviews focus on one type of grid-edge resource only, for example, on residential thermal energy storage [85] or deferrable loads [86]. Others focus on specific segments of the electricity grid, such as on the residential context [30, 87], or on microgrids [88–91]. Previous reviews have analysed particular coordination methods such as market-based coordination [81–83, 92–96], optimisation [97], particle swarm optimisation and genetic algorithms [98] and RL [9]. Others have investigated specific technological tools enabling coordination, such as distributed ledger technologies (DLT) [99] and digital tools such as modelling, simulation and hierarchical control [100]. Tohidi et al. review grid-edge resources coordination indirectly, by investigating the possible interactions between local and central electricity markets [101]. Guerrero et al. investigate the technical issues associated with implementing behind-the-meter DERs coordination strategies in

a low-voltage network [102]. Despite all these individual thematic reviews, there was no systematic review and taxonomy of the field of coordination of resources at the edge of the electricity grid with applicability across energy technologies. The terminology used to categorise DER coordination strategies within the literature has overlaps and inconsistencies, with terms such as "peer-to-peer", "multi-agent" or "transactive energy" which may refer to coordination frameworks with fundamentally different structural features. The resulting unproductive linguistic ambiguity impedes effective communication and understanding in the field, which can hinder both academic progress and collaboration with industry.

This section bridges this gap to bring greater clarity to the classification and terminology in the field of DER coordination. The principal contributions of this section are:

- The development of a novel comprehensive taxonomy for distributed resources coordination strategies, which aims at clarifying the structural features of coordination strategies, as the terminology currently used to describe them is often ambiguous.

- The identification of key research themes corresponding to the coordination categories. A systematic review of 84,741 publications relevant to the coordination of distributed resources at the edge of the electricity grid was performed using a structured topic search query.

- The analysis of 93 coordination strategies mapped to the taxonomy through a detailed literature review to illustrate the wealth of coordination strategies corresponding to each category of the taxonomy.

These contributions will help identify the key gaps and promising areas for this thesis to explore.

The rest of this section is organised as follows. In Section 2.1.1, a novel synthesised taxonomy of coordination strategies is developed based on the types of agency, information flow structure and game type. The relevance of this

categorisation is analysed in key associated research trends identified by a systematic literature review. Section 2.1.3 demonstrates the ambiguity of some of the current terminology used in the field, which is not consistently aligned with the structural features of coordination categories, and highlights how the taxonomy helps resolve this ambiguity. It then provides a more detailed review to clarify the boundaries of the coordination categories by mapping 93 selected recent coordination strategies onto the taxonomy. Finally, Section 2.1.4 looks at potential applications of the taxonomy for coordination strategy selection. It is argued that context-specific heterogeneous complementary strategies are needed to coordinate different flexibility sources in energy systems and suggest specific paths forward for research, such as residential energy coordination.

2.1.1 Taxonomy of distributed energy resources coordination strategies

This subsection develops a novel taxonomy for the categories of DER coordination based on objective structural features of the individual strategies.

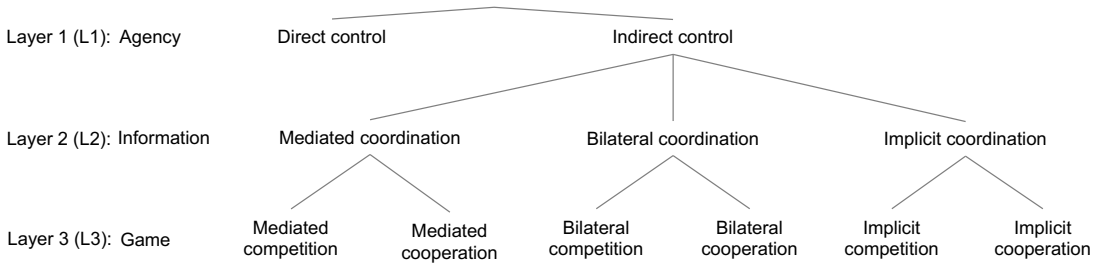


Figure 2.2: Systematic taxonomy of categories of coordination, based on the answer to three questions corresponding to the three layers of the classification: (L1) Agency: Are coordinated units operated independently? (L2) Information: How is individual information shared? (L3) Game type: Do units compete or cooperate?

As shown in Figure 2.2, three questions corresponding to the three layers of the taxonomy L1, L2 and L3 are used to systematically classify any coordination strategy into exhaustive and mutually exclusive structural categories: (L1) Agency: Are coordinated units operated independently? (L2) Information: How is individual information shared? (L3) Game type: Do units compete or cooperate?

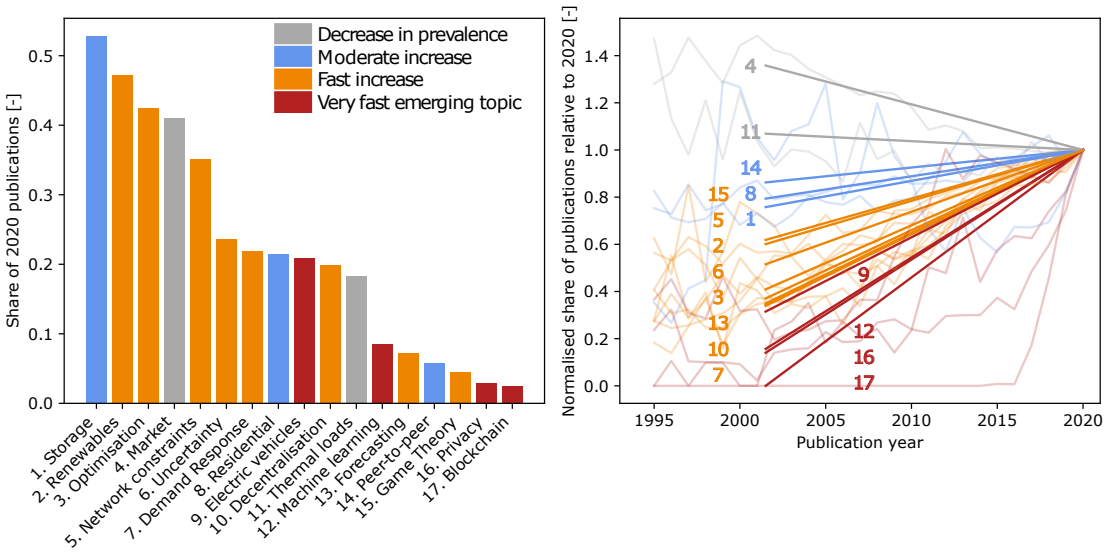


Figure 2.3: Research theme prevalence in the field of grid-edge resources coordination. The left-hand side plot shows the prevalence of research themes in 2020, i.e. the share of all publications on the coordination of grid-edge electricity resources that include theme-related keywords in their titles and abstracts. Faint lines on the right-hand side show the prevalence of themes over time normalised by that in 2020. Opaque lines show the change between the average normalised prevalences of themes in the first half of the period (1995-2007) and that in 2020. Moderate, fast, and very fast increases correspond to average prevalences in the first half of less than 100%, two thirds and one third of that in 2020, respectively. See Appendix A.2 for the details of the theme identification. Note that all themes have increased in absolute numbers over this period, as the total scholarship under study has increased; this figure analyses the relative importance of individual themes within the scholarship.

The rest of this section presents these three elucidating questions in more detail and illustrates their relevance in key associated research themes identified in a systematic literature review of grid-edge energy resource coordination. A structured topic search query in the Scopus search engine [103] was conducted, selecting literature that lies at the intersection of the concepts of coordination, grid-edge participation and electric resources (see Appendix A). Key research themes were identified within this body of literature using keywords from titles and abstracts. The absolute and relative prevalence of these major themes are displayed on Figure 2.3, so that trends corresponding to the main layers of the taxonomy may be identified.

Layer 1: Agency. Are coordinated units operated independently?

Direct control refers to the case where a central entity has full access to the information from all units and can decide on their control actions. A system objective is pursued regardless of whether this is beneficial for single units. In *indirect control*, prosumers make decisions at the local level.

This thesis defines a unit as one or multiple energy resources providing flexibility in operation in the same location, whose assets and information are operated by the same entity. The proposed taxonomy is agnostic to the type of coordinated unit, and each structural category is applicable across a wide range of applications. Given the increasing urgency of climate change and consequently the need for electricity decarbonisation, there has been a rapid increase in interest in the coordination of both renewable energy assets (47.2% prevalence in the literature on grid-edge energy resource coordination in 2020 in Figure 2.3) and of key associated enabling flexibility resources. Integrating large shares of intermittent renewable electricity generation will require flexibility to align the power generated with consumption and provide operational services [104]. Words relating to storage thus occurred in just over half (52.7%) of the relevant literature. While storage traditionally takes the form of standalone stationary batteries, a very fast-emerging number of publications are investigating the coordination of EV batteries. Moreover, other elements of flexibility may also be conceptualised as virtual storage, such as that provided by demand response [105, 106], which is also a fast-emerging research theme. Coordinating thermal loads has for example long been studied in this body of literature, though there is now a slight relative decline in interest. On the other hand, residential flexibility resources receive growing attention due to the outstanding DSR potential provided by the increasing electrification of household loads and ownership of smart resources. However, there are significant challenges to their coordination due to the large number of small units and the required interaction with user needs, which prevent direct monitoring and controllability of distributed units.

Given these trends, the question of agency is critical, as the direct control of resources from different owners with different objectives and resources is challenging

[15]. As a result, interest in research themes that place independent preferences and decisions at the heart of coordination strategies, such as peer-to-peer trading, game theory and blockchain structures, has been increasing, though they are so far only mentioned in 5.7%, 4.5% and 2.5% of the relevant literature respectively (see Figure 2.3). The choice of *direct* or *indirect control* depends on the level of intelligence and flexibility of individual units, the alignment of individual interests, as well as other contextual legal and physical constraints [88].

Layer 2: Information. How is individual information shared?

Indirect control strategies can, in turn, be subdivided between *mediated* coordination where a central entity collects information[1] about prosumers, *bilateral* coordination where prosumers only communicate information bilaterally with one another, and *implicit coordination* where personal information is not shared, with at most one-way communication of market information to prosumers without feedback of personal information.

The information structure of coordination strategies is increasingly relevant as interest in the decentralisation of energy resources is rapidly increasing in the literature (see Figure 2.3), transforming the way data is owned and communicated. Recently developed strategies have sought to complement the existing centralised control by fully utilising not only the generation and flexibility offered locally but also the distributed data ownership, computation power and communication capabilities. While exchanging information provides value, it also raises numerous ethical, trust and very fast increasing privacy concerns (see Figure 2.3). Data gathered may be improperly used, including both information directly collected from prosumers and other sensitive information about users' habits and lifestyles inferred from their behaviour and interactions [50]. As such, a trade-off exists between the value of sharing information and both the degree of privacy [52] and the costs of communication and control infrastructure.

[1]Information refers to load and generation curve predictions, bids or constraints, for example.

With decentralisation and limited local data availability comes increasing uncertainty, which is a fast-increasing research theme (23.5% prevalence in 2020), especially as the share of intermittent and unpredictable electricity-generating technologies increases. When both physical resources and data ownership are distributed, complete and instantaneous information flows for determining adequate controls cannot be obtained. An optimal control signal based on inaccurate information yields suboptimal outcomes, with potential negative impacts on physical systems such as the electricity grid. While some techniques such as robust and least-regret optimisation account for this uncertainty, the fields of forecasting (7.2% prevalence in this corpus in 2020 in Figure 2.3) and machine learning (ML) (8.5%) are of fast and very fast increasing interest respectively. Depending on the information structure, these can help bridge a lack of data at the local level. Forecasting aims at reducing uncertainty in predictions, although renewable resources availability and behaviour-influenced demand are inherently unpredictable, especially at the local scale. In ML, agents learn incrementally from collected experience and can take statistically optimal actions with incomplete information within an uncertain, stochastic environment. The use of ML for DER coordination is facilitated by the increasing availability of data measurement in the electricity grid and by reduced computational requirements compared to physical-based models [100].

Data availability and communication are therefore key in determining which type of coordination strategy to use.

Layer 3: Game type. Do units compete or cooperate?

Finally, the coordination strategies are classified based on the type of game prosumers are playing.

Prosumers may willingly *cooperate* towards the maximisation of common global objectives, such that some additional social value may be obtained relative to the sum of individual utility if decisions were made in isolation [76]. In certain situations, agents could perform actions with sizeable benefits for society as a whole, but would derive insufficient personal benefits from bringing these positive

externalities in a competitive framework [106]. Therefore, regulatory intervention and cooperation between various actors are needed to unlock these benefits. A common objective may simply be the equitable achievement of all private self-interested objectives, or may be broader in scope. A common objective may for example be managing network constraints. Although the challenge of maintaining grid operation within acceptable physical limits has always been prevalent in distributed electricity resources coordination, the share of publications dealing with power distribution networks management has been steadily increasing due to the aforementioned decentralisation and uncertainty of resources (Figure 2.3) [83]. Concerns include voltage fluctuations and imbalance, current harmonics, network congestion and stability issues [100], especially in the distribution grid where independent assets are not readily controlled, and at the interface between transmission and distribution grid. Failing to account for such network constraints in coordination strategies may lead to negative impacts during operation and increased investment costs [102].

Alternatively, prosumers may *compete* with one another. They seek to maximise their own local utility[2] only, whether purely based on profits or taking into account personal preferences. This market-based control aligns loosely with definitions of "transactive energy" with "users [...] considered as self-interested with heterogenous preferences" [102].

Note that although individual prosumers aim to maximise their own utility only, the overall market mechanism may be designed to direct those self-interested decisions towards additional objectives such as network management or aggregator profits, to obtain collective outcomes as close as possible to an optimisation result [14]. Electricity markets particularly differ from markets in other sectors in that they need a designer and a system operator, who may solve an optimisation problem

[2]Utility was defined by Jeremy Bentham as "that property in any object, whereby it tends to produce benefit, advantage, pleasure, good, or happiness (all this in the present case comes to the same thing) or (what comes again to the same thing) to prevent the happening of mischief, pain, evil, or unhappiness to the party whose interest is considered" [107]. A utility function, in turn, is an economist's convenient representation of an individual's preferences that permits mathematical analysis [108].

for optimal dispatch. Therefore, the result of trading can be modelled as an optimisation problem, although no optimisation actually occurs. Ideal market signals are analogous to dual variables in optimisation problems for marginal pricing frameworks [109]. As an example, long-term climate change mitigation may be incentivised using marginal carbon costs pricing signals to align personal interests to global ones [110]. However, without the inclusion of externalities in prices (greenhouse gas emissions, network constraints management, grid losses, network utilisation), markets may not maximise the welfare of the entire system. Optimal real-time operation of energy resources is therefore an interdisciplinary matter, where both optimisation and market-based approaches can co-exist and complement each other – they were each mentioned in just above 40% of the selected body of literature in 2020 (Figure 2.3). Market-based control of electricity assets has been extensively researched concurrently with the wave of liberalisation of the system in the 1990s, while the research focus is now increasingly on optimisation strategies as the reach of information and communication technologies (ICT) is extending to the edge of the grid (Figure 2.3).

2.1.2 The ambiguous terminology of control architectures

Using the taxonomy developed in Section 2.1.1, it is found that the terminology used in the literature does not elucidate clear structural features of coordination strategies. Linguistic ambiguity in the field hinders the clarity of structural assumptions about control paradigms, with terms being used to denote inconsistent meanings.

We take as examples the terms “multi-agent”, “peer-to-peer” and “transactive energy”. As shown in Table 2.1 and the bullet points below, each of these labels has been given to such varied fundamental approaches to coordination that they have lost their specificity.

- Strategies have been labelled as “multi-agent systems” in contexts often departing from the established definition proposed by Wooldridge [111]: “An agent is a computer system that is capable of independent action on behalf of its user or owner. In other words, an agent can figure out for itself what

it needs to do in order to satisfy its design objectives, rather than having to be told explicitly what to do at any given moment. A multi-agent system consists of a number of agents, which interact with one another, typically by exchanging messages through some computer network infrastructure". For example, a framework was described as a "multi-agent system" although distributed units had no agency, and residential agents were directly controlled centrally [112]. Table 2.1 show the label is further used all across the landscape of coordination strategies, from direct control to implicit coordination. While there is no right or wrong definition, the various meanings tied to the term have prevented this labelling from conveying clear specific meaning on its own.

- Similarly, Table 2.2 illustrates the widely varying levels of specificity with which "peer-to-peer" systems are defined. Moreover, publications both describe P2P mechanisms as inherently cooperative [33, 76] and competitive [113].

- While most frameworks labelled as "transactive energy" refer to competitive frameworks, this is inconsistently used. Thus, in one source, transactive energy systems were defined as using "automated device bidding [...] through market processes" [114], while in another, distributed optimisation techniques have been framed at transactive energy, with all agents cooperating to reach a single objective by decomposing the problem at the user level [102].

Coordination category	"multi-agent" labelling	"peer-to-peer" labelling	"transactive energy" labelling
Direct control	112		102, 115
Mediated competition	116–120	96, 121–123	102, 115, 124–126
Mediated cooperation	127–129	33, 130	37
Bilateral competition		96, 121, 131, 132 67, 133, 134	102, 126, 133
Bilateral cooperation	89, 135	33, 113	102
Implicit competition			102, 114, 115, 126
Implicit cooperation	136, 137		

Table 2.1: Example uses of the labels "multi-agent", "peer-to-peer" and "transactive energy" in different structural coordination categories across the proposed taxonomy. The terminology therefore loses specificity.

Reference	Peer-to-peer definition
134, 138	The trading between suppliers and consumers
102	Users trading energy among themselves with limited or no intervention of a third party
83, 123, 139	The sharing of resources at the edge of the network, at the distribution level
121, 140	Selling and buying of surplus electricity for small-scale residential and commercial producers and consumers
133	The coordination of a large number of small-scale producing and consuming units
96, 131	The ability of individual prosumers to make independent decisions, being able to choose when to trade, how much, and at what price
125, 141	Markets allowing prosumers to engage in bilateral trades
67, 142	Mechanisms allowing prosumers to negotiate with one another directly
113	Mechanisms allowing prosumers to negotiate with one another directly and without any direct influence of a central controller
132	A network in which members share resources and information to attain energy-related objectives, without any intervention from a third party controller, and where any peer can be added or removed without altering the operational structure of the system

Table 2.2: In increasing order of specificity, examples of definitions for "peer-to-peer" trading found in the literature. The label therefore no longer points to a specific structural coordination category.

The wealth of concepts and definitions appearing in the literature testifies to the

rapidly increasing interest the field is attracting. However, overlapping terminology may be counter-productive as we lose the ability to label individual coordination structures clearly and precisely. The taxonomy developed in Section 2.1.1 may help resolve this ambiguity and improve communication by providing objective structural criteria for classification. As the body of literature concerned with coordinating resources at the edge of the electricity grid is growing exponentially and extends to every layer of our electricity systems, precise terminology is needed for relevant problems to begin to be understood and assessed.

2.1.3 Detailed literature review

This section conducts a detailed review of the literature on the coordination of grid-edge electricity resources through the lens of the taxonomy developed in Section 2.1.1.

This section seeks to resolve the ambiguity pointed to in Section 2.1.2 by demonstrating how all strategies can be mapped onto the taxonomy categories, using a sample of 93 coordination strategy definitions from the literature. This illustrates the wealth of strategies corresponding to each of these categories and clarify their structural differences and similarities.

2.1.3.1 Direct Control

Direct control is identified in the top layer (L1) of the taxonomy in Figure 2.2. Units forfeit their data and control to a central entity. Different terms in the literature align with this category of control architecture, such as "centralised control" [89], "direct load control" [8], and "centralised dispatch" [67].

Direct control may be technically accomplished in different ways. Receding horizon global optimisation frameworks are commonly proposed [143–146], where the uncertainty of optimisation inputs can be accounted for in with stochastic optimisation [144] and by subjecting the scheduler to the worst-case PV predictions [146] among others. The amount of data shared by units to allow the centralised entity to make control decisions may vary from "top-down switching" where the entity uses statistics to directly turn on and off loads without consumer information

given, to "centralized optimization" where the central entity truly can optimise the search space thanks to full access to information [115]. Design and operation of the system may be considered together under a total system or bi-level optimisation [147, 148]. Schedulers using RL are proposed that directly control residential appliances [149]. Direct control may also be rule-based, for example, to synchronise fridges' thermal storages [150] or to design some direct-load control architecture for an aggregator based on priority stacks [151] or other heuristics [112]. In event-based direct control, customers receive incentive payments for allowing the utility a degree of control over certain equipment; the utility can reduce the loads in response to various trigger conditions such as grid conditions or system temperature [84].

Direct control methods are particularly suited to small-scale microgrids [143], or if the aggregator owns resources directly. However, individual domestic households may not be inclined to give up control of appliances in their own homes, expressing privacy and security concerns. Direct control may moreover pose a significant computation burden or even be computationally intractable at large scale [122]. It could necessitate extensive ICT infrastructure and be highly vulnerable to the single point-of-failure of the central controller or communication links [83, 88, 100]. Leveraging full information availability and control, typical objectives for direct control strategies are global operation cost minimisations, energy arbitrage, peak shaving, load shifting and ancillary services (see Table A.3).

2.1.3.2 Indirect control

In *indirect control* (Figure 2.2: L1), coordinated units have agency and are operated independently at the local level. This can be either a manual or automated response. Indirect control may be broken down into *mediated*, *bilateral* and *implicit coordination* (Figure 2.2: L2) as presented below.

Mediated coordination In *mediated coordination*, a central entity collects information about prosumers (Figure 2.2: L2) and redistributes information back to them, such as matchings between peers, price signals or partial results from a central optimisation for use in local computation. Two-way communication

allows the utility to monitor real-time flexibility availability and leads reliably to intended outcomes as information is available to steer individual actions in the desired direction. This category aligns with the concept of "coordinated approach" [102]. Different well-established mathematical theories, such as auction theory and optimisation theory, can be used to design mediated coordination.

This approach is most suited to situations with infrastructure and acceptance for reliably and safely sharing individual information with a central entity. Nevertheless, prosumers may have security and privacy concerns over sharing their information centrally. Biased information, due to the inaccuracy of forecasts or to strategic behaviour [142], may moreover lead to inefficient or even infeasible decisions. It is not guaranteed that prosumers will have the will, interest, or ability to receive, interpret, and respond according to the centrally computed signals [42]. Indeed, both local ICT infrastructure and prosumer engagement are necessary to utilise flexibility.

In the third layer of the taxonomy (Figure 2.2: L3), mediated coordination can be broken down into *competitive* and *cooperative* categories as defined below.

- *Mediated competition* involves competitive individual units which maximise their own objective function only. The mediator collects information from units and sends back signals they believe will incentivise globally optimal action, such as price signals or prosumer matchings. Although participating units are selfish, further objectives may be served depending on the design of the strategy, such as the aggregated provision of ancillary services, the minimisation of global operational costs and the flattening of load curves (see Table A.3).

  For example, the central entity may send unidirectional price signals to customers based on information such as prosumers' costs, constraints and day-ahead forecasts, corresponding to a proposed definition of a "community market" [96]. In a "price-responsive mode", prosumers make decisions in response to market price signals by the operator [126]. Pricing signals corresponding to Distribution Locational Marginal Pricing (DLMP) with

import-export price spreads can incentivise adequate prosumer decisions [67]. Pricing signals that are adaptive to centrally monitored consumption levels internalise the externality effect of customers, where an increase in customer demand would create a negative externality effect for other customers' price rates [30]. Given knowledge of previously negotiated prosumer trades, the central distributor can compute and allocate losses using its global knowledge of the distribution network and transactions [124]. RL can inform both the dynamic price signal [120, 152], and the prosumer response to price signals [120, 153]. The exchange of information between prosumers and the mediator may also be iterative. In such a non-cooperative Stackelberg game, both utilities and consumers try to maximise their utility by iteratively updating their prices and demand, respectively [154]. In another example, prosumers learn to compute a Nash equilibrium through iterations of exchange of information between prosumers and operator, with electricity prices dependent upon the aggregate demands at each time step [155].

Alternatively, another approach to mediated competition uses "organized markets" [84] analogous to wholesale markets, where the central entity collects bids, sets trades and matches prosumers centrally. The labels "system-centric matching" [121], "coordinated market" [96], "transactive control" [115] and "explicit demand response" [98] fit in this category. Instead of sending control signals directly, bidding approaches promote price discovery, seeking to reveal the marginal value of consumption given fragmented unknown private technical characteristics, personal preferences and opportunity costs of participants. This bidding reveals the information required to obtain control signals without directly using private information such as supply and demand functions used in formal optimisation [114]. Examples of this include unilateral auction mechanisms [123], continuous double auction mechanisms [102, 114, 122], demand reduction bids where customers send bids consisting of their available demand reduction capacity and their requested price [84], an auction mechanism where the difference between accepted offers and accepted

bids are allocated for congestion costs [156], and nested negotiations between DSO and aggregators and between aggregators and prosumers [125]. RL algorithms may be used to facilitate bidding either to refine individual bidding strategies [117, 119, 140, 157, 158] or to dictate the double auction market clearing [35, 116].

- In *mediated cooperation*, the mediator collects information from units (Figure 2.2: L2) and sends back signals to the prosumers, expecting them to cooperate (L3). The aim is to maximise system-wide welfare by minimising total costs, sometimes against immediate individual preferences or interests. Customers may need technical and financial support to install the necessary infrastructure to react accordingly to signals from the utility [84]. The mediator may take on a "social planner" role, a benevolent coordinator who chooses an economic policy either to maximise a social welfare function or to attain a Pareto efficient allocation [159]. Typical applications of mediated cooperation seek to address both prosumer utility maximisation and cooperative objectives such as peak shaving, ancillary services provision or minimisation or operation costs (see Table A.3). Though distributed units work together to achieve common goals in both cases, mediated cooperation differs from direct control in that the mediator does not have full access to personal information for central direct control of units. Instead, only partial personal information is shared so that central signals may guide cooperation without full knowledge of individual utility functions or device states, while final computations and decision-making are on the individual level.

  The mediator's role is to collect and redistribute information from and to prosumers so that they can take actions compatible with global objectives. In a semiautonomous mode of operation [160], instructions on how to respond to locally available grid signals are updated centrally and regularly based on power systems characteristics and sent to local appliances. In an iterative method, the power system operator shares network information with prosumers, who

can then privately arrange trades, considering their marginal losses and constraints, and communicate updated trades to the operator [161]. Based on the new set of trades, the operator then broadcasts updated network information. In other examples of iterations between central and distributed computations, prosumers run local optimisations with augmented terms to align local prosumer coordination with global constraints and goals, and send back results to adapt iteratively centrally computed signals informing, in turn, the local optimisations [138, 162].

In other strategies, prosumers cooperate with the help of a mediator by forming coalitions to achieve their goals. In a "community-based market", members agree on common goals and trade together as a coherent group in markets, the trades being handled by a community manager [33]. This is similar to the "energy collective" where a supervisory node facilitates the interface to different markets [37]. Prosumer coalitions can be computed centrally using game theory based on their information, and the optimal operation is solved centrally for the coalition [76]. In another example, prosumers participate in a coalition formation game with an auctioneer acting as an intermediate, i.e. a mediated competition process precedes the cooperation [130].

RL algorithms have been proposed where agents interacting with a central entity learn to cooperate to pursue common goals. Agents can use RL with individual agent fitness functions that include the objective functions of other agents while satisfying local operation constraints, and learnings are shared within cooperative swarms to update their knowledge rapidly [118]. In other frameworks, the central entity sends each prosumer information about eight neighbours [127, 128]. Each prosumer learns to pursue common goals for the nine prosumers, such as aligning loads to renewable energy provision and avoiding high total load. In another strategy, an aggregator collects information on the flexibility of building agents and the distribution network [129]. Each building energy management system (EMS) then aims to meet flexibility requests by the aggregator with minimum discomfort, using a

multi-agent Q-learning method that includes an explicit model of the other agents.

Bilateral coordination In *bilateral coordination*, prosumers only communicate information bilaterally with one another with no central authority or organised pool (Figure 2.2: L2). This category is most suited to contexts where communication infrastructure is available, and privacy requirements are compatible with the bilateral communication of individual information. Robustness to communication failures increases relative to mediated coordination as there is no longer a centralisation of data with a single point of failure. As the system size increases, the number of communication iterations until algorithm convergence increases, requiring adequate computational resources with limited communication network latency for feasibility [102].

The safe way of implementing distributed transactions to ensure data protection is an ongoing subject of research. Many trial projects use DLTs such as the blockchain, allowing for a decentralised marketplace without an intermediary. An incorruptible, secure digital ledger records financial transactions permanently in a decentralised fashion [163], in multiple points without a single point of failure [91]. However, they present legal, environmental and implementation risks. Legal risks of non-compliance to regulations such as the European Union General Data Protection Regulation (GDPR) arise, as personal data cannot be erased and stakeholders cannot control their own data, with no clear accountable entity [164]. Furthermore, negative externalities associated with blockchains have been well documented, due to proof-of-work processes requiring ever greater computational power and energy consumption for mining, causing environmental and health impacts contradicting the very aim of decarbonisation [165]. Finally, implementation risks such as scalability concerns, facilitated money laundering, fraud, and tax evasion should be carefully weighed before creating a dependency at a large scale of our electricity systems on blockchain technologies [166].

We now further differentiate the *bilateral competition* and *cooperation* (Figure 2.2: L3) categories.

- In a *bilateral competition* framework, peers directly negotiate energy transactions with one another (Figure 2.2: L2) and choose trades to maximise their utility (see Table A.3). This is analogous to "peer-centric" mechanisms [121], "decentralised markets" [96], and "bilateral transactive bids" [126]. The offer of autonomy, expression of individual preferences and market transparency may appeal to prosumers. However, agents may exhibit bounded rationality, unable to know the competitors' production decisions and profit functions [167, 168] and to process all the necessary information to perform a cognitively burdensome exhaustive optimisation of rational utility in real decision-making situations. They may often rely on simple rules of thumb or decision heuristics, which may widen the optimality gap [169] and dampen the reaction of agents to market or policy signals [170]. To avoid potentially serious consequences for the physical grid system, markets should therefore be designed to elucidate consumer preferences and appraise their impact on both market outcomes and the physical power networks [33, 168].

  Some bilateral competition strategies have been proposed with blockchain-managed bidding mechanisms [131–133]. Others rely on bilateral contract networks, multi-sided matching markets where units form bilateral downstream and upstream contracts to sell outputs and buy inputs [102]. The potential for stability of a bilateral contract network was extensively analysed by Fleiner et al. [171] and further explored by Morstyn et al., who established a fully distributed, iterative P2P negotiation mechanism [67, 134] .

- In *bilateral cooperation*, prosumers also use bilateral negotiation, but to cooperatively implement theoretically-proven market solutions for the whole system (Figure 2.2: L3). Prosumers compute local responses, considering not only their personal preferences and profits but also system objectives and constraints. A range of cooperative objectives may be pursued (see

Table A.3). This category closely aligns with the definition of "cooperative control of multi-agent systems" [172], "multi-agent control strategy" [89], and "distributed control" [100].

Bilateral cooperation strategies can be implemented using iterative negotiation to reach shared objectives [67, 102]. In other proposals, common objective functions are maximised using variations of distributed optimisation such as dual decomposition and the alternating direction method of multipliers (ADMM) [33, 173]. For example, distributed optimisation can be used [113], where global and local constraints are decoupled by applying Lagrangian multipliers. Notwithstanding, scalability is a concern for these methods. In ADMM, although there are established proofs of convergence, the convergence time may be sensitive to problem-specific numerical properties and may be operationally impractical [37]. In other proposed strategies, agents use transfer learning with distributed W-learning to achieve local and system objectives [135].

Bilateral cooperation may be vulnerable to the risk of strategic behaviour or gaming by market participants, as the convergence to optimal cooperative outcomes depends on information supplied by individual units and on their cooperating rather than trying to maximise their individual utilities only [142]. Cooperative game theory-based profit-sharing allocations that incentivise cooperation in a way that is robust to strategic behaviour are exponentially complex, limiting scalability [76]. Moreover, computational issues may arise as the complexity increases with the number of DERs at scale [67]. Safeguarding measures include ring-fencing the distribution networks, with a clear definition and allocation of distribution costs in incentive regulation [174]. This approach is suitable for systems with larger numbers of nodes and high complexity, and offers cost reductions for systems that frequently need to be expanded relative to centrally controlled systems with more expensive communication and control infrastructure [100].

Implicit coordination In *implicit coordination*, prosumers do not share personal information with a central entity or their peers. However, they may monitor their current wider information environment (e.g. wholesale prices) or utilise information about past system characteristics to inform their independent decision-making (Figure 2.2: L2). While bilateral trades are effective coordination tools, the physical reality of the electricity network only recognises injection and extraction points, with electrons flowing independently of the bilateral financial transactions the market participants agree upon. The problem of coordination thus boils down to the sum of individual decisions to import and export electricity. This case corresponds to the definition of "decentralised control" [100] where "the control decisions are made individually at each DER by its local controller using the local information".

Advantages of implicit coordination include reduced complexity and costs of the ICT infrastructure, enhanced privacy, self-control and acceptability for users, robustness against failures, adaptability to changing environments and reliability [100, 102, 175]. Beyond the necessity of using communication-less multi-agent learning due to privacy and cost constraints, independent and autonomous agents can thus also lead to more robust and general-purpose systems, limiting the impact of communication delays, unreliable information (voluntary or not) or failure of other agents [175]. The simplicity of implicit coordination reduces implementation failure risks due to the lack of interoperability between different smart grid elements, as connectivity between individual hardware and software components was reported as the most common obstacle reported in real-world programmes [15, 84]. One-way communicating devices are highly cost-effective as they incur lower upfront costs than two-way communication, though they do not allow monitoring and verification of the DSR impact with accurate precision [84]. A suboptimality gap will exist as the strategies cannot be informed by real-time personal data from all units, and due to unexpected problems in the case of uncooperative operation [100]. Therefore, these strategies are most suited where the cost of ICT infrastructure for each unit outweighs the potential benefits of communication, particularly for small units with privacy concerns.

We now map coordination strategies onto the *implicit competition* and *implicit cooperation* coordination categories.

- In essence, *implicit competition* is the status quo scenario. Most consumers today do not share their data and behave competitively (Figure 2.2: L3), i.e. solely maximise their own utility to serve local goals such as individual costs and comfort (see Table A.3). This paradigm is also called "price-reactive system" [115], "smart pricing" [84], "autonomous mode" [126] and "uncoordinated approach" [102]. Critical-peak pricing, time-of-use pricing (TOU) and real-time pricing are common ways of implementing implicit coordination [84].

  As self-interested units are only concerned with the individual scale, many strategies leave the realm of distributed units coordination to focus on single-unit local energy management, for example, with load scheduling algorithms under TOU pricing [176]. RL models can help optimise personal utility by exploiting opportunities for energy arbitrage [177, 178], learning optimal appliance scheduling decisions by interacting with user feedback [179], and conserving energy while ensuring user comfort [180]. Rule-based and RL-based control of heating, ventilation, and air conditioning (HVAC) resources have been proposed [181]. The uses of particle swarm optimisation and genetic algorithm have also been investigated to control thermal energy storage [98].

  Self-interested responses to price signals, without factoring in the impact on the whole system of the sum of their individual actions, result in suboptimal system outcomes, especially if deployed at a large scale. For example, a concern is that all loads receive the same incentive, the natural diversity on which the grid relies may be diminished [182], and the peak potentially merely displaced, with overloads on upstream transformers. A case study thus results in capacity issues for high DER penetration if global network constraints and objectives are not taken into account in decision-making [102]. Moreover, uncoordinated reactions to price signals may be difficult to predict without knowledge

of devices' states and end users' preferences [115]. Contrary to mediated competition, where the coordination signals are dynamically updated based on real-time communication of individual information, in implicit competition, the prices are sent unidirectionally and thus need to be carefully selected ahead of operation. The market mechanism should be designed such that self-interested schedules add up to limit suboptimality.

- In the *implicit cooperation* category, units do not share information (Figure 2.2: L2) but make individual decisions cooperatively to optimise global objectives (L3).

  Cooperating without either direct, centralised control or bilateral sharing of personal information has been proposed to reach system-wide benefits statistically. Prosumers seek to statistically assess the impact of their actions on achieving common objectives. Within this category are found the "decentralised control strategies" [89] and "autonomous control" [100], which focus on voltage and droop control [90], a line frequency control method based on local information only. Fully autonomous modes of control have been defined to offer a hard-wired primary response to grid frequency deviations [160].

  This thesis makes the argument that the space of possible control and coordination strategies and methods that would fall under this categorisation is under-researched, with very few strategies exploring implicit coordination beyond frequency control (see Table A.3). Two such examples were identified [136, 137], where agents aim to minimise charging peaks and local voltage deviations, respectively. In the latter, agents maximise a local, personal reward which penalises local voltage deviations, rather than a global one. It is thus formally an example of competitive multi-agent reinforcement learning (individual rewards), though this coordination is classified as cooperative due to the nature of the reward (helping to manage the grid).

The mapping of coordination strategies onto the proposed taxonomy presented throughout this section is synthesised in Table 2.3 and extended in Table A.3.

These tables illustrate both the wealth of possible strategies within each category and the lack of specificity of the terminology which is used across different structural categories.

In Table A.3, the type of objectives pursued is classified to aid in selecting adequate strategies given contextual structural constraints and objectives. As presented in this section, multiple direct control strategies provide load shifting, peak shaving and ancillary services, while competitive strategies tend to focus on individual prosumer utility. Typical cooperative strategies seek to minimise operation costs while considering prosumer utility. Strategies that consider network losses and constraints are also identified, as network management was previously identified as a critical challenge for the coordination of DERs. Note that individual citations may be listed multiple times if multiple strategies were defined in a paper. Quoted descriptions for each strategy are included, which show that the labelling of the strategy in the paper alone does not allow the reader to understand which paradigm it falls under unambiguously. Where specified, Table A.3 lists individual units that are coordinated in the papers (local generation, storage, thermal control, generic flexible load), which shows that the strategy classification is agnostic to the unit type it is applied to. Strategies that use reinforcement learning are also identified, illustrating how this tool may be used under various paradigms, as each paradigm may be implemented in numerous ways beyond the control, communication and incentive structure.

Coordination category	Example keywords for strategy description	Refs.
Direct control	"centralized", "centralised dispatch", "direct load control", "event-based DSR", "model predictive control", "Operator instructions", "optimal power flow", "optimization", "top-down switching"	67, 84, 89, 100, 102, 112, 115, 126, 143–147, 149–151
Mediated competition	"adaptive consumption-level pricing scheme", "bidding strategy", "coordinated market", "demand response", "double auction", "dynamic pricing algorithm", "incentive-based demand response", "indirect customer-to-customer energy trading", "market-based control", "non-cooperative", "organized markets", "price-responsive", "P2P market","Stackelberg game", "transactive control", "transactive energy", "unidirectional pricing"	30, 35, 67, 84, 96, 98, 102, 114–117, 119–126, 132, 140, 152–158
Mediated cooperation	"community-based", "coordinated multilateral trades", "distributed demand response", "distributed optimization", "distribution locational marginal costs and hierarchical decomposition", "energy collectives", "joint action learning", "prosumer coalitions", "P2P trading", "transfer learning"	33, 37, 76, 118, 127, 128, 130, 138, 160–162
Bilateral competition	"auction", "bilateral contracts", "bilateral transactive bids", "blockchain", "decentralized market", "distributed", "local energy market", "matching theory", "P2P trading"	67, 96, 102, 130, 131, 133, 134, 171
Bilateral cooperation	"bilateral energy trading", "decentralized", "distributed","distributed dispatch", "parallel", "P2P"	33, 67, 89, 100, 102, 113, 135
Implicit competition	"autonomous", "energy arbitrage", "implicit demand response", "load scheduling", "markets", "price DSR", "price-reactive systems", "rate-based", "uncoordinated"	84, 98, 102, 126, 176–181
Implicit cooperation	"autonomous control", "decentralised", "fully autonomous", "multi-agent"	89, 100, 136, 137, 160

Table 2.3: Summary of the detailed literature review of distributed energy resources coordination strategies mapped onto the taxonomy categories. For an extended version of this table, including a breakdown of each strategy, see Table A.3. Single citations may be listed under multiple categories if multiple strategies are defined in a paper. Each coordination category may be labelled by a variety of terms, while individual terms may be used to describe strategies across different categories. This reflects both the variety of possible strategies within each category and the lack of specific terminology, which leads to ambiguity.

2.1.4 Application of the taxonomy for coordination strategy selection

This literature review has developed a taxonomy to map any coordination strategy using non-ambiguous, objective structural classification criteria.

This section illustrates the application of this taxonomy to select adequate types of coordination strategies based on the context of their application, such as the one in this thesis. Section 2.1.4.1 argues that, to achieve deep decarbonisation, an ecosystem of control architectures suited to each specific context in all layers of the energy system is needed. Particularly, it highlights the need to identify and focus on under-integrated and under-researched niches. Section 2.1.4.2 then illustrates this argument using the taxonomy to highlight the potential of implicit cooperation for coordinating residential energy, whose flexibility is so far under-utilised.

2.1.4.1 Selecting complementary coordination strategies for deep decarbonisation

Deep decarbonisation requires the coordination of a heterogeneous system of interlinked technologies, infrastructures, markets, regulations and user practices, such that every carbon-emitting activity contributes to efforts, beyond the low-hanging fruits [183]. In the context of energy systems, it involves unlocking the contributions to environmental and social welfare from assets at all levels. This includes both the traditional centralised generation plant at the transmission level and the heterogeneous distribution-network flexible loads beyond the reach of centralised strategies. Energy systems involve a plurality of stakeholders (e.g. business model developers, smart home systems manufacturers, individual generators and consumers, regulators, flexible loads aggregators) deploying a wide range of technologies that require control (e.g. microgrids, residential energy, industry, buildings, distributed renewable generation, transport fleet charging).

To identify appropriate DER coordination strategies for each context, researchers can borrow established heuristics to investigate where to focus research to maximise impact with limited resources.

An obvious place to start to assess a methodology is: "Is this the most effective thing you can do?" [184]. One size does not fit all; the choice of coordination category from the taxonomy in this paper will depend on individual coordination contexts and aims. Structurally differing coordination paradigms with specific advantages and drawbacks for different applications have been presented in Section 2.1.1. For example, while it may be worth investing large amounts in communication and control infrastructure to exploit the flexibility potential of larger industrial assets fully, obtaining complete access to information and control to perform real-time centralised optimisation may be impractical or not most effective where the value of flexibility is smaller in individual distributed assets. The trade-offs between the additional value enabled by communication and the costs of implementing it, both in terms of equipment needed and consumer acceptance, must be particularly appraised. Other important criteria to be considered in individual contexts include the level of intelligence and flexibility of individual units, the type of prosumer (residential, commercial, industrial, military), stakeholder involvement, underlying motivation and drivers, supporting infrastructure required, integration with existing tariff structure, legal requirements, unit physical features (location, ownership, size), network assets response time, devices and communication failures, security risks, electrical grid conditions and the desired robustness of the system [83, 84, 88, 114]. Coordination strategy-specific risks, such as those related to privacy and security, are mitigated using multiple types of interoperable strategies suited to different challenges. Therefore, a heterogeneous mix of coordination strategies across the space of possible coordination categories is needed to go from shallow utilisation to deep integration of each remaining niche whose flexibility is so far underutilised.

Another important question may be asked regarding setting priorities for developing coordination strategies: "Is this area neglected?" [184]. We have shown that research is increasing exponentially in the field. There is a need to frame research relative to existing scholarship to avoid replicating efforts and focus on under-explored areas. Certain resources and tools have historically been extensively researched, such as the central optimisation of large power plants and industrial

users. There is value in identifying under-integrated and under-researched segments with locked-in flexibility and value, such as in the residential sector.

2.1.4.2 Potential path forward for future research

The residential sector is now considered as a broadly recognised remaining niche with so far under-integrated flexibility. This example illustrates the use of the taxonomy to select adequate coordination strategies in a given context.

As stated in Section 1.1.2, residential sites have a pivotal role to play in helping facilitate the integration of renewable energy generation, which represents significant untapped opportunities.

Identifying the specific remaining hurdles for the coordination of residential energy flexibility, suitable coordination strategies can be identified using the taxonomy proposed in this paper. Firstly, computational and social acceptance limitations mean that direct control would not be best suited at scale. There may be difficulties in computing central solutions due to scalability issues for large numbers of independent units, especially as residential customers are not motivated to invest substantial amounts in managing their electricity usage. Moreover, accurately forecasting individual residential consumption and generation profiles for day-ahead optimisation is challenging [84]. Secondly, due to both ICT infrastructure cost constraints as the residential flexibility potential is broken down into many small households with potential individual benefits too low to justify the cost of communication links, and to privacy concerns, personal information may not be readily shared in real time.

Implicit cooperation, which keeps personal information at the local level while encouraging cooperation towards global objectives, is an under-researched coordination category that would be particularly suited to unlocking the potential systemic value of residential energy flexibility. As the communication and computation burden can become unwieldy in large-scale systems, pre-learned policies computed using RL and implemented in a decentralised fashion may help coordinate the so-far largely untapped residential energy flexibility. It avoids issues of costs, privacy

and technical risks associated with both centralised and bilateral communication by allowing consumers to cooperate without sharing their data. Incentives could be designed to incorporate social objectives in the operation of distributed home energy systems without excessive interference in personal comfort and utility.

A significant hurdle for the development and implementation of such data-driven methods is the availability of large datasets on EV consumption and at-home availability, PV generation, and household consumption for training and testing data. Data quality determines the results of data-driven methods such as ML predictions or RL policies [185], and should be as much of a focus as the algorithmic development. Therefore, novel methods are required to meet the needs of both large-scale datasets and the inclusion of real-life patterns. This is further discussed in Section 2.3.1.

2.2 Multi-agent reinforcement learning: approaches and challenges

Section 2.1 has identified the RL-based implicit coordination of residential energy flexibility as a promising under-explored avenue for coordinating residential DERs. This section of the literature review describes the multi-agent reinforcement learning research landscape as a promising methodology for this application.

First, this section commences in Section 2.2.1 by comparing reinforcement learning to other control frameworks. Then, Section 2.2.2 presents key aspects of reinforcement learning methodologies, which serve as the foundational element of MARL. Then, the landscape of MARL architectures is presented in Section 2.2.3.

2.2.1 Motivation for reinforcement learning control framework selection

This subsection starts by comparing potential control frameworks, namely optimisation, control theory, and reinforcement learning, given a global objective function to be maximised (or minimised) and a set of constraints.

2.2.1.1 Optimisation approaches

Convex optimisation is a well-established field with strong theoretical underpinnings that have been researched for over a century and can be applied to problems known ahead of time to be convex [66]. The framework aims to select decision variables x that minimise an objective function $f_0(x)$ given constraints $f_i(x) \leq b_i,\ i = 1, ..., m$, where the objective and constraint functions are convex, which means they satisfy the inequality

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y) \tag{2.1}$$

for all $x, y \in \mathbb{R}^n$ and all $\alpha, \beta \in \mathbb{R}$ with $\alpha \geq 0,\ \beta \geq 0$. and where the equality constraints are affine.

As discussed in the first part of this literature review, despite these solid theoretical underpinnings and decades of collective experience in their computational implementation, using centralised convex optimisation is not achievable in practice for the decentralised, computationally scalable coordination of DERs without private data for the following four reasons.

1. Firstly, centralised optimisations are hindered by a lack of centralised data availability. In the standard convex optimisation paradigm, the global problem is solved centrally for all decision variables, and standard optimisation methods cannot be used without full knowledge of the system's inputs and dynamics [186]. The use of optimisation techniques may result in privacy breaches, as confidential parameters such as transmission details and individual bids may be inferred from the optimal outcome [187]. This may deteriorate trust and enable strategic market actors. However, if access to data is limited, either due to these privacy concerns or to limited communication capability, the solvers face a knowledge problem, especially in complex and stochastic environments. As such, in residential energy, agents only have *partial observability* of the system due to both the stochasticity and uncertainty of environment variables such as individual residential consumption and generation profiles, and to the

privacy and infrastructure cost constraints that hinder communication between agents during implementation [188]. Optimisation with inaccurate data or model would lead to suboptimal outcomes: the problem of finding a global optimum *given that global information is available* differs from the problem of finding a robust policy *given that only partial information is available*. Not relying on shared information may also improve the robustness of the solutions to failure of other agents, communication delays, and unreliable information, and improve adaptability to changing environments [175].

2. Moreover, even if privacy and acceptability were not a concern, prosumers may not have the capacity to predict their consumption and generation accurately in order to perform day-ahead optimisation – as DSR moves from large-scale industrial to distributed residential customers, processes become more complex, uncertain and unpredictable.

3. Thirdly, even if full and perfect local data were available, centralised optimisations would be intractable at the scale of millions of homes [189]. The large number of units may pose a computational challenge for direct, centralised, real-time control. Scalability has thus been identified as the main challenge for consumer-centric markets [37].

4. Finally, even if we had perfect data and computational power, it may not be possible to obtain a convex model to represent the real-life complex electricity grid environment. Firstly, it may have sources of non-convexity. While simplifications can be made to render the problem convex, real-life results may differ from the simulation results. Moreover, every building, organisation or technology has different flexibility potential, presents its own challenges and has different combinations of flexibility sources [43]. Due to the heterogeneity of users and behaviours needing different parameters and models, the large-scale use of model-based controllers is cumbersome [190].

These issues could be partially mitigated by using model predictive control (MPC) methods. MPC has a mature control theory and can achieve high control performance by solving a real-time optimisation problem at each step that minimises the objective function over a truncated time horizon [191]. At each stage, the decisions based on this lookahead of a small number of stages are implemented for a short time into the future, before the optimisation for future time steps is repeated based on updated information [192]. MPC mitigates the increasing uncertainty for further time steps and reduces the computation by truncating the time horizon. Nevertheless, its implementation in practical settings is still in its early stages due to factors such as its complex design and demanding online computational requirements [193]. Moreover, MPC still requires a sufficiently accurate model of the environment dynamics, and performs online estimation of the cost function, which usually leads to a high online computation complexity that may require powerful advanced hardware for real-time applications [193]. This increased computational cost is also no guarantee of superior performance relative to RL – in a building energy system experiment, RL had an equivalent or superior performance to MPC [193].

In terms of the privacy concerns of centralised convex optimisations, various variations on the standard centralised computation have been proposed. In [187], privacy-preserving random perturbation strategies are adopted, by combining stochastic (chance-constrained) programming and differential privacy. Alternatively, one could consider a distributed optimisation approach to that problem, with local computations and iterative exchange of partial solutions. However, as discussed in Section 2.1.3.2, this iterative approach based on dual price variable adjustment (e.g. dual decomposition, ADMM) poses challenges, such as the number of iterations until convergence. The infrastructure and computational cost may be too large, while not relying on shared information may also improve the robustness of the solutions (see Section 2.1.3.2).

In summary, optimisation-based techniques are based on solid theoretical foundations but have slow computation speed at scale and are not directly applicable in a scalable, multi-agent, fully distributed framework. However, despite the

limitations of standard convex optimisation techniques for real-time, large-scale implementation, this thesis considers that those techniques may help during the algorithm development and learning phases. Optimisations will be used both by providing an upper bound for the achievable objective function, and to provide information to the learners during the simulated learning phase.

2.2.1.2 Control theory

RL and control communities often seek to tackle similar dynamical system planning and control problems, although the communities have stayed vastly disjoint [186]. According to Recht, "control is the theory of designing complex actions from well-specified models, while reinforcement learning often makes intricate, model-free predictions from data alone".

Adaptive optimal-control typically are *indirect* methods, in which controls are recomputed from an estimated system model at each step [194]. In simple model-based schemes, where the reward and dynamics functions follow well-studied characteristics, proven methods may be used successfully. Given a complete and accurate model of the Markov decision process (MDP), an optimal control rule can thus be found by applying one of several dynamic programming (DP) algorithms. Standard linear quadratic regulator (LQR) problems will have analytical solutions to minimise costs using standard linear algebraic techniques [186]. However, The LQR optimal response is not robust to noise, disturbance or modelling imprecisions that are a primary feature of the fully decentralised coordination of residential energy flexibility in the electricity grid. Moreover, model estimates are inherently complex, making adaptive methods, in which the optimal controls are estimated directly, more attractive. Model-free reinforcement learning may thus require significantly less computation at each time step than indirect adaptive optimal control methods using conventional DP algorithms [194].

Control theory and RL are not diametrically opposed, but rather, the insights they offer overlap and can be synthesised. According to Sutton [194], "[a]lthough its roots are in theories of animal learning developed by experimental psychologists,

reinforcement learning has strong connections to theoretically justified methods for direct adaptive optimal control". Reinforcement learning methods can be conceptualised as a computationally simple, direct approach to the adaptive optimal control of non-linear systems, as a synthesis of dynamic programming, stochastic approximation methods and control theory [194].

2.2.1.3 Reinforcement learning approaches

The three main reasons for using RL in this research over other control frameworks are to maintain a model-free approach, to allow for coordination under partial observability, and to avoid computational issues for large-scale real-time implementation.

Firstly, a model-free RL approach does not require a priori knowledge and may be more suited for the context under study in this thesis. Its data-driven approach allows it to look into more complex real-life applications that may lie beyond the reach of accurate models [186]. Energy end-users are heterogeneous, with different patterns of behaviour, each potentially needing different model parameters and even different models. Therefore, the large-scale implementation of model-based controllers becomes cumbersome [190]. RL is based on frequentist statistics, directly interacting with an uncertain environment without necessarily modelling it explicitly [195]. This search for optimal control given unknown dynamics [186] is useful both in case of partial ignorance of the task, of partial understanding of the complex interactions between variables or parameters [189], and for non-convex environments and non-smooth reward designs. This model-free, black-box approach however also lacks theory support – as a data-based, empirical approach rather than an exact algorithmic approach, this strategy calls for experimentation to ascertain the performance and robustness of different variations.

Secondly, RL allows for coordination under partial observability. Central optimisations need accurate global data and direct control abilities. RL can overcome the constraints of centralised convex optimisation for residential energy coordination by allowing for decentralised and model-free decision-making based on uncertain and partial knowledge. Agents can only use local data to maximise

the statistical expectancy of global rewards in a distributed fashion. This is more cost-efficient and privacy-preserving, as well as improving reliability, robustness (in the case of communication issues with one or multiple agents or of inaccurate data) and computational efficiency. Instead of producing optimal results using inherently inaccurate data, RL more realistically searches for sequential decisions which statistically maximise rewards under uncertainty [186]. At each time step, the uncertainty from future electricity consumption, transport requirements, PV production and other agents' behaviour is implicitly taken into account.

Finally, RL offers a way to avoid computational issues for real-time implementation at a large scale. Once the RL policies are learned, no further computations or communications between agents are needed for real-time, fast and decentralised implementation of the pre-trained policies. The implementation time does not increase for increasing numbers of agents, as decisions are made in parallel. Even if agents had a perfect knowledge of the problem and could perfectly control all the variables, the amount of computational resources needed to perform centralised, real-time optimisation may drive prices up or simply be infeasible. RL can overcome the constraints of centralised convex optimisation for residential energy coordination by allowing for decentralised and model-free decision-making based on partial knowledge. Approximate learning methods such as reinforcement learning may be more computationally scalable and more efficient in exploring high-dimensional state spaces relative to computationally heavy exact global optimisation methods [98, 189].

2.2.2 Reinforcement learning

RL is first defined in Section 2.2.2.1. Then, categories of RL methodologies are listed in Sections 2.2.2.2 to 2.2.2.5. These subsections intend to introduce a selection of concepts used in this thesis, rather than to be exhaustive.

2.2.2.1 Definition

RL is an artificial intelligence (AI) framework that considers sequential decision-making by goal-oriented agents who learn by interacting with an environment

[185]. *Agent* refers to both the learner and the decision maker[3]. The agent interacts with an *environment*, which comprises everything outside itself. This terminology deviates from that typically used in engineering and the field of control theory, as mapped in Table 2.4.

Reinforcement learning	Control Systems
Agent	Controller
Environment	Controlled system, plant
Action	Control signal

Table 2.4: Mapping between the terminology used in the reinforcement learning and engineering fields [185].

RL is used to solve MDPs, a classical formalisation of sequential decision-making, where actions influence not just immediate rewards but also subsequent situations, or states, and, through those, future rewards [185]. At time step $t \in \mathcal{T}$, an agent is in a state $s^t \in \mathcal{S}$ and selects an action $a^t \in \mathcal{A}$ according to the agent's policy π. The environment then produces a reward $r^t \in \mathcal{R}$, and agents transition to a state s^{t+1}. MDPs involve delayed rewards and, as such, the need to trade off immediate and delayed rewards.

Reinforcement learning is one of the three fundamental types of machine learning:

- Supervised learning, which is most commonly researched in the field of machine learning, whereby learning is informed by a training set of labelled examples provided by a knowledgeable external supervisor.

- Unsupervised learning, which is typically about finding structure hidden in collections of unlabelled data.

- Reinforcement learning, where agents seek to maximise a reward signal without labelled examples of correct behaviour. Reinforcement learning is driven by analogues of reward and punishment in neurophysiological mechanisms – reinforcement learning, across species, involves dopamine [196].

[3]In the context of this thesis, agents are independent computer systems acting on behalf of prosumers

As an increasing wealth of data is collected in local electricity systems, RL is of growing interest for the real-time coordination of DERs [9, 197]. As classified in Section 2.1, numerous RL-based coordination methods have been proposed in the literature for residential energy coordination. Scalability and privacy protection limitations however remain.

2.2.2.2 Tabular methods and deep learning

In tabular methods, the states and actions spaces are restricted enough to represent value functions as tables. The simplest and most popular such algorithm is Q-learning [198], a value-based, off-policy temporal difference control method (see Section 4.2.1). Advantages of using this method include its simplicity, interpretability, and general proof of convergence to the optimal value function, which is available under the conditions that the state-action pairs are discrete, and all state-action pairs are repeatedly sampled (sufficient exploration) [199]. However, these conditions are often inapplicable for high-dimensional and possibly continuous state-action spaces (Bellman's "curse of dimensionality"). This proof of convergence is also lost in the case of multi-agent RL considered in this thesis. Moreover, table-based representations of models and value functions become intractable for large state and action spaces [200].

In contrast, deep reinforcement learning (DRL) uses Artificial Neural Networks (ANNs) of interconnected simple processing units, computational models inspired by biological nervous systems [197]. Neural networks comprise multiple processing layers, each trained to minimise the empirical error between targets and estimates. There is no fundamental conceptual difference between deep and non-deep reinforcement learning, apart from the use of neural networks for function approximation [186]. Approximate methods are most useful in problems with high dimensional, arbitrarily large and possibly continuous state spaces, as is the case in most problems approaching real-world complexity [197]. The network starts with random weights and gradually adapts itself to maximise rewards in a given task [196]. Most ANN learning is based on the *fire together, wire together* rule, stated in the 1940s by

the neuropsychologist Donald Hebb [201]. Hebbian learning strengthens often-used connections. When two linked units are activated simultaneously, the weights are adjusted to make this more likely in future. As explained in [196]:

> The algorithm assumes that the error in an output unit is due to error(s) in the units connected to it. Working backwards through the system, it attributes a specific amount of error to each unit in the first hidden layer, depending on the connection weight between it and the output unit. [...] Proportional weight changes are then made to the connections between the hidden layer and the *preceding* layer.

Contrary to the Q-tables, to learn a parameterised estimate of the value function or the policy, neural networks do not require an exponential increase of data when adding extra dimensions to the state or action space. In addition, deep RL has been shown to be able to learn different levels of abstractions from data in complicated tasks, even with low prior knowledge [195]. However, tabular theoretical convergence guarantees are lost when switching to parameterised representations such as neural networks.

Given their respective advantages, both tabular and deep learning methods will be investigated in this thesis.

2.2.2.3 On- and off-policy

RL agents aim to learn an optimal behaviour, or *target policy*, using experience collected by the *exploration policy* used to make decisions.

In *on-policy* methods, the exploration policy is equal to the target policy. This introduces a bias when used with a replay buffer, as the previous trajectories are usually not obtained solely under the current updated policy π. Therefore, a main drawback of these methodologies is the inefficient use of data, with the current data sample discarded after every update.

Off-policy algorithms, on the other hand, learn from an exploration policy that differs from the target policy. This is more powerful and general, and allows the agents to learn from previously generated data, such as that generated by experts. It is thus more data efficient as the agent can learn in batches from previously collected data even if the data was collected with an outdated policy that no longer

matches the current estimate of the optimal policy. Sample efficiency can further be improved by using *experience replay*, allowing us to reuse samples for future policy evaluation steps. However, off-policy methods are more complex, have higher variance, and are slower to converge [185].

2.2.2.4 Model-based and model-free RL

Model-based reinforcement learning relies on a *model* of the environment, where the model refers to transition functions for the state transitions and reward function for state-action pairs. This model can be either explicitly given or learned from experience. Once a model is available, planning can be done through value iteration, lookahead searches through potential action trajectories or other trajectory optimisation controls. A model-based approach thus requires using a planning algorithm, which may be computationally expensive [195, 200].

In comparison, model-free RL algorithms do not use such explicit transition probability distribution of the MDP. The optimal policy is estimated without using or estimating the dynamics.

Whether to adopt model-based or model-free approaches depends on the ease with which a reliable model of the environment can be obtained. In some instances, it is easier to learn a model of the environment due to the particular structure of the task (less complex or with more regularity). In others, it may be more efficient to estimate the policy or the value function directly. François-Lavet proposes the following analogy: in a labyrinth, it is clear how actions affect states and rewards and the model dynamics may easily be generalised, although the successful policy may be intricate. In a busy road crossing on the other hand, high stochasticity renders a model-based approach difficult, although one could easily learn a straightforward policy (e.g. moving forward except if an obstacle appears) [195].

This thesis uses model-free RL, as it deals with a problem that is more similar to the latter, with high stochasticity of the non-deterministic environment and partial observability of individual agents that do not have all the information to

run a planning algorithm, but where relatively straightforward policies on local actions may be learned that yield overall satisfactory savings.

2.2.2.5 Value, policy, and actor-critic reinforcement learning

As model-free methods, value-based, policy-based and actor-critic algorithms do not make use of any model of the environment [195]. The agents' experience can be used to directly update an estimate of the optimal value function or policy [200].

Value-based RL implicitly finds the optimal policy by finding the optimal value function; this provides sample efficiency and learning stability and excels in sample efficiency. This thesis will start by exploring this route in Chapter 4.

On the other hand, in policy-based RL, the optimal policy is computed by manipulating the policy itself directly. This allows for faster convergence as it directly optimises the quantity of interest, and remains stable under function approximation. Policy-based learning is also effective in high-dimensional, continuous and stochastic continuous action spaces and learning stochastic policies. However, this requires on-policy learning, with high gradient variance and sample inefficiency.

In actor-critic methods, both policy and value functions play an important role, and separate independent memory structures are used to represent either. The policy structure, i.e. the actor, selects actions, and the estimated value function, i.e. the critic, criticises the actions made by the actor [185]. Thus, actor-critic methods aim to exploit advantages from both value- and policy-based methods. The learning can be off-policy and, in the policy evaluation step, the value functions can be updated using a more stable estimate of the value function from several trajectories through the environment, rather than the high variance in single experience samples. This also means that the influence of samples extends beyond a single update, improving data efficiency. This thesis will therefore use an actor-critic methodology in Chapter 5.

2.2.3 Multi-agent reinforcement learning architectures

This subsection now lays out the differences between single and multi-agent RL in Section 2.2.3.1. Then, it presents the fundamental MARL architectures for the centralised training of decentralised policies in Section 2.2.3.2. This will provide the basis for methodology selection in the remainder of this thesis.

2.2.3.1 From single to multi-agent reinforcement learning

This thesis concerns itself with a system, i.e. a group of autonomous, interacting entities sharing a common environment [202]. While in RL, a single goal-oriented agent learns sequential decision-making by interacting with an environment [185], in MARL, multiple agents simultaneously learn by exploring a shared environment.

Several real-life problems involve interaction between multiple agents, where emergent behaviour and complexity arise from agents co-evolving together. The ability to effectively scale RL to scenarios featuring multiple agents is vital for the development of productive AI systems [203]. As indicated by Boden [196]

> [AI agents] use their highly limited intelligence in cooperation with – or anyway, alongside – others to produce results that they couldn't achieve alone. The interaction between agents is as important as the individuals themselves.

In the wake of single-agent RL's successes, MARL has therefore gained rapid traction in recent years, as plotted in Figure 2.4. The number of yearly publications mentioning "reinforcement learning" and "multi-agent reinforcement learning" in their title, abstract and keywords has thus been increasing exponentially by 22.0% and 38.0% per year between 2000 and 2022, respectively. MARL can help address problems of real-world complexity and shows great potential for resolving diverse cooperative multi-agent issues, including the coordination of robot swarms [204] and autonomous vehicles [205]. It also opens benefits such as the decentralisation of the learning task in the presence of communication or computation limitations, robustness in implementation (if one agent fails, the others can keep coordinating) and scalability [202].

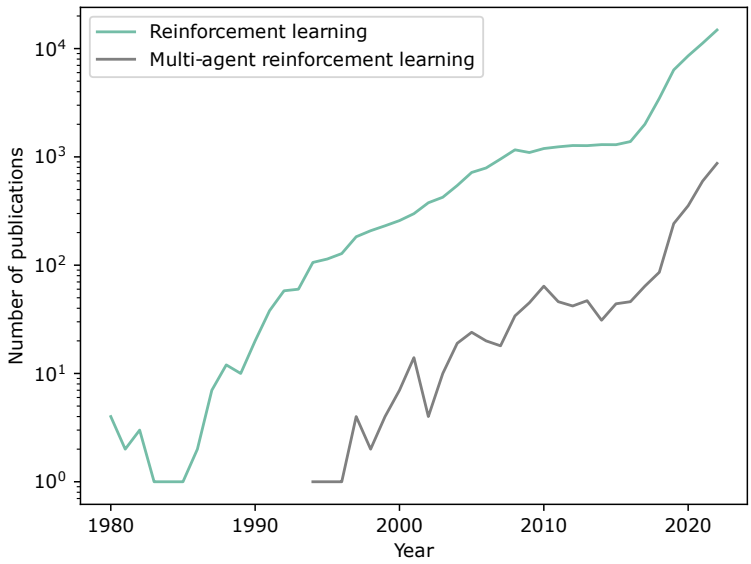


Figure 2.4: Number of publications in Scopus which mention either 'multi-agent reinforcement learning' or 'reinforcement learning' in their titles, abstracts and keywords.

Notwithstanding, traditional RL methods such as Q-learning or policy gradient were initially developed for single agents, and their performance drops for increasing numbers of agents under partial observability in stochastic environments. The environment stationarity assumption in single-agent RL is violated as each agent faces a moving-target learning problem due to the other simultaneously learning agents [206]. This invalidates the convergence properties of most single-agent RL algorithms. Moreover, the scalability of algorithms, already problematic in single-agent RL, becomes an even greater cause for concern [202]. Methods such as policy gradients usually exhibit very high variance when the coordination of multiple agents is required, which means learning becomes more unstable [203]. Moreover, while proofs of convergence can be derived in the case of multi-agent reinforcement learning in deterministic environments [207], this does not hold for stochastic environments.

Further methodological innovation is therefore needed for multiple such agents to learn to cooperate. While there is no guarantee of finding the optimal policies in the realistic, stochastic environment under study in this thesis, the aim is to find policies that maximise the rewards over different cases with variability between runs.

In addition to the problem of harmonisation between the agents, agents are faced with the difficulty of incomplete information with respect to the choice of action [207]. While most existing AI-based DSR research thus assumes fully observable tasks [197], in MARL, different agents may have access to different information. As such, the direct controllability of the energy resources considered in this thesis is challenging, as they are owned by different owners with different objectives and resources and are subject to privacy, comfort and security considerations [15]. Developing policies that maximise returns under uncertainty can also be advantageous, as it improves coordination robustness in a stochastic environment. Reducing the reliance on shared information and increasing the flexibility of online problem-solving can improve adaptability to changing environments [175]. This robustness to changing environment conditions will be demonstrated in Section 6.3.

Transitioning from single-agent to multi-agent RL also raises the question of whether agents adopt a cooperative approach, working to maximise a shared reward, or a competitive approach, where they strive to maximise their individual rewards. In the competitive case, each agent has its own reward function it seeks to maximise. While finding policies that maximise the summed discounted rewards for all agents is unattainable, the objective is to achieve an equilibrium point where no agent can enhance its reward by altering its policy, provided that all other agents maintain their policies. In cooperative multi-agent MDPs (MAMPDs), on the other hand, all agents use the same reward function. Hence, it is possible to determine policies that denote not only equilibrium points but also optimal policies that result in maximum discounted reward for all agents [207]. This is the case in this thesis, which investigates a fully cooperative problem with a common return to be maximised.

2.2.3.2 Centralised training with decentralised execution

In MARL terms, the implicit residential energy flexibility coordination framework considered in this thesis corresponds to a cooperative partially observable Markov decision process (POMDP) with decentralised execution of policies. Due to partial observability or communication constraints, each agent must thus learn a

decentralised policy conditioned only on local observations to maximise a shared objective function. However, the decentrally implemented policies can themselves be trained in a centralised or decentralised manner. In the *centralised training with decentralised execution* (CTDE) [208] paradigm, the training is centralised in a simulated environment, with access to additional information about the environment (e.g. global state) and other agents. This centralisation of information during training is a powerful tool for gaining a full understanding of the system and learning to coordinate during the rehearsal phase. There are different ways of implementing this paradigm.

Table 2.5 summarises common classes of cooperative MARL approaches used to solve a Dec-POMDP, which are presented below.

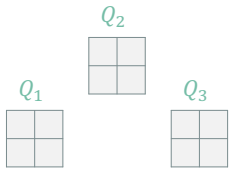
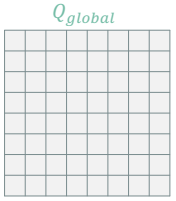
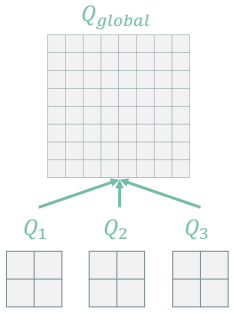
MARL Approach	Schematic illustration
Independent learning [209–213]	
Centralised multi-agent policy gradient [203, 214–216]	
Value function factorisation [217–222]	

Table 2.5: Three common classes of multi-agent reinforcement learning approaches to solve a decentralised partially observable Markov decision process (Dec-POMDP)[4].

Independent learning The simplest type is *independent learning*, where each agent treats other agents as part of the environment and learns independently. Agents concurrently learn individual value estimators whilst still using a global reward definition. For instance, in independent Q-learning (IQL) [209], each agent i learns a *decentralised* action-value function Q_i based only on individual observations and actions. Numerous extensions to this approach exist, such as Deep Q Networks (DQN) [210]. Research has been conducted on replay stabilisation methods [212] (case studies with up to 5 agents) and task specialisation with transfer learning [211] (case studies with up to 3 agents), among others. Independent Q-learning can perform satisfactorily at scale in some multi-agent tasks, where partial observability does not constrain the learning of successful policies [223]. As a general rule, however, cooperative independent learning faces challenges to cooperation as it cannot explicitly represent interactions between agents. Among others, independent learners may converge to local optima that are incompatible with the global optimal (*Pareto selection problem*), individual contributions to global rewards cannot be distinguished from the stochasticity of the environment and the behaviour of other concurrently exploring agents (*stochasticity* and *alter-exploration* problems), and the environment is *nonstationary* due to the evolving policies of agents during learning, violating the MDP stationarity assumption [136, 206]. This thesis starts by investigating this option in Chapter 4, as the most straightforward approach most amenable to interpreting the coordination mechanism requirements between agents in a given task.

Centralised multi-agent policy gradient In contrast, another category of CTDE algorithms is the *centralised multi-agent policy gradient* method [203, 214–216], in which each agent learns a centralised critic with a decentralised actor. The *centralised and monolithic* critic estimates the global value of all possible combinations of states and actions of all agents, which are only available during

[3]Note that the value functions are represented as tables (as would be the case in a Q-learning implementation) for illustration purposes, though these can be estimated using function approximators such as neural networks.

centralised training, to estimate the *centralised* action-value function Q_{tot}. Compared to Q_i, Q_{tot} improves coordination by capturing the interdependent effects of all agents' actions and guiding the optimisation of decentralised policies. This makes use of all available information and can handle coordination problems. Yet, learning a satisfactory estimate of Q_{tot} can be impractical since it directly conditions on the global state and joint action, which can grow exponentially with the number of agents. The approach is therefore not practical for more than a handful of agents, running into exponential computational time requirements and memory problems. Performing separate policy gradients for each agent using the centralised critic, assuming others are fixed, can also lead to sub-optimal policies, in which no single agent wishes to change its action unilaterally [222]. Moreover, extracting individual decentralised policies from a global critic is not straightforward.

Value function factorisation *Value function factorisation* methods [217–221] constitute a third category of CTDE algorithms. The centralised action-value function Q_{tot} is not directly learned but instead represented as a mixing function of individual action-value functions Q_i to enable easy decentralisation and enhance scalability. Value-based MARL algorithms first proposed to take the global estimator as the sum of individual value functions (Value-Decomposition Networks, VDN) [224], then in QMIX [217] to use a richer class of action-value functions that factorise the global value into individual functions via non-linear monotonous functions in a mixing network. Value decomposition has been shown to be an effective approach in most environments and shares the major advantages of centralised training, especially in environments with dense rewards [223], such as in DER coordination. The only requirement is that the arguments of the global value function maxima are equivalent to the arguments of a set of maxima on individual value functions. Subsequent algorithms such as QR-MIX [220] and RMIX [221] have extended this principle to include finer risk sensitivity. However, how to best represent and learn Q_{tot} is still an open question.

The Factored Multi-Agent Centralised Policy Gradients (FACMAC) algorithm [222] incorporates elements of both the centralised multi-agent policy gradient and the value function factorisation paradigms. It uses centralised but factored value networks in an actor-critic framework, to reap benefits from both the value factorisation and the actor-critic superiority relative to pure value-based learning in a number of tasks. This improved coordination allows agents to learn more coordinated behaviour and escape sub-optimal solutions. Chapter 5 of this thesis therefore builds on this methodology.

2.3 Research gaps

We identify the following research gaps, at the interface of the DER coordination and MARL fields.

2.3.1 Local energy system environment for RL algorithm testing

The literature on novel CTDE MARL architectures predominantly uses virtual "toy" benchmark game environments, designed explicitly for evaluating and developing MARL algorithms. As such, the StarCraft multi-agent challenge (SMAC) [225] has emerged as a prominent benchmark in the field of multi-agent reinforcement learning (MARL) since its introduction in 2019. However, the widespread adoption of SMAC has given rise to various concerns, including problems associated with environment overfitting and selective reporting of results, which raise doubts about the reliability and validity of research findings when using it as a benchmark [226]. If MARL is to leave the laboratory setting, benchmark environments must tackle real-world application problems such as DER coordination.

Therefore, this thesis aims to provide a residential DER coordination environment freely as a testing environment for MARL algorithm benchmarking on a realistic, complex problem. The proposed environment goes beyond the CityLearn local energy system environment, which has been used for RL testing [227], not only

by including EV availability modelling but also by integrating a training and testing data generator.

While Figure 2.3 has shown that machine learning is a very fast emerging topic in the field of DER coordination, a significant hurdle for the development and implementation of data-driven methods is indeed the availability of large datasets on EV consumption and at-home availability, PV generation, and household consumption for training and testing data. Beyond the need for a DER coordination training and testing model, there is a need for an integrated data generation tool. Indeed, while large amounts of residential energy data are recorded, training directly on available data is often unsatisfactory given:

1. The privacy and cost constraints of data collection, or cost of access to datasets that are not freely available without a licence or privileged access. Obtaining energy data frequently poses a significant challenge for the development of energy communities [228]. This can result in substantial time and financial resources being expended. Generally, open-access databases offer rather restricted access to comprehensive energy consumption and production profiles, as the establishment of open-access data initiatives is fraught with numerous legal and occasionally ethical obstacles and inquiries. Companies may be wary about sharing their energy data outside their business [228].

2. The limited number of years of data collection available (e.g. for electric cars, for which we predominantly have smart trial data from early adopters), or the limited number of subsequent days of data available for a given household, which hinders consistent simulation of a home for more extended periods. For example, the National Travel Survey offers at most a week of travel data for a given household [229].

3. The labour-intensiveness of pre-processing data, with efforts repeated across individual projects, as datasets are often not in a usable format, or not self-consistent across different days. Data quality has been identified as a challenge for the adoption of AI in the smart energy industry [230]. A major hurdle

identified by energy community initiators is thus that of data formatting standards and the quality of the acquired data [228].

While agent-based modelling approaches have been previously adopted to model residential data such as EV patterns [231], training data should reflect real-life resource intermittency and behaviour variability to minimise training losses robustly without over-fitting. Purely synthetic data often lacks these characteristics. Moreover, bottom-up models such as CREST [232] rely on assumptions on dwelling activities and thermal-electrical demand modes for generating data.

Therefore, novel methods are required to meet the needs of both large-scale datasets and the inclusion of real-life patterns. A standard residential energy data generation tool that could interface with a local energy system benchmarking environment to generate continuous daily energy data for several days in a consistent manner, both in terms of profile magnitude and behavioural clusters, would greatly benefit the research community. While [233] first proposed the use of generative adversarial networks to generate smart grid-related data, a more extensive tool is required to generate UK data, integrate this tool directly into a MARL benchmarking framework, and include EV data generation.

2.3.2 Multi-agent reinforcement learning for fully decentralised implicit cooperation

While there is a growing interest in research on both DER coordination and MARL, bridges must be built to identify new and appropriate MARL methodologies that are best suited for the scalable coordination of residential energy flexibility without sharing private data. The detailed literature review of DER coordination has revealed that RL is used only in a minority of proposed implicit coordination strategies. When systematically examining the literature on grid-edge DER coordination more generally, only 1.55% of the titles and abstracts mentioned reinforcement learning, and only 0.14% multi-agent reinforcement learning[4].

[4]using the search terms in Appendix A.1, as of the 14th April 2023

RL-based implicit coordination strategies could allow prosumers to coordinate their loads cooperatively in a decentralised manner. Particularly, given the challenges of coordination between agents in partially observable MDPs discussed in Section 2.2.3, MARL cooperation mechanisms suited to the problem of DER coordination should be identified or developed.

However, as discussed in Section 2.1.3.2, most RL-based DER coordination strategies take a competitive approach, which could adversely affect energy users and the grid. Full implicit cooperation, which keeps personal information at the local level while cooperating towards a global reward or objective function, has been thus far under-researched beyond frequency control. Moreover, the applicability in more complex scenarios with residential electric vehicles and smart heating load scheduling problems has not been considered. Finally, in existing proposals, the convergence slows down for an increasing number of agents, and scalability beyond eight agents has not been investigated. Scalability, in particular, is therefore a central concern in MARL [202] as previous MARL algorithms have not scaled to the number of agents required at the scale of a feeder in DER coordination.

The present thesis aims to bridge this gap and enhance the feasibility of MARL algorithms for coordinating residential energy flexibility in practical applications.

2.3.3 An assessment of the impact of residential energy implicit coordination on distribution networks

Finally, the potential impact of MARL-based implicit coordination on the electricity grid and for end energy users has not yet been evaluated in the literature. While there are numerous studies aiming to assess the impact of DER coordination approaches on the electricity grid, namely the impact of direct control [67, 102, 143–146], and of mediated [67, 102, 121–125, 129, 161] and bilateral [102, 113, 133, 234] coordination, only one reference in the literature review assessed the impact of RL-based implicit coordination strategies on network constraints with individual agent rewards [137], and none in a full MARL-based implicit cooperative approach

with a global reward to be maximised jointly. An assessment of the potential impacts of this novel approach is therefore needed.

2.3.4 Bridging the research gaps

The first research gap resented in Section 2.3.1 will be addressed as part of the first research question in Chapter 3: *Can the efficacy of algorithms that coordinate residential energy flexibility be assessed?* The research gap identified in Section 2.3.2 is then solved in two steps over chapters Chapters 4 and 5 to answer the two research questions related to the performance and computational scalability of coordination algorithms: *Can residential energy flexibility be successfully coordinated without sharing private data? Can this be achieved in a computationally scalable manner?* Finally, the research gap laid out in Section 2.3.3 corresponds to the fourth sub-question tackled in this thesis and will be investigated in Chapter 6: *Can the algorithms achieve positive impacts for energy users and the grid?*

All models are wrong, but some are useful

— George E. P. Box

3

A local energy system environment for use in reinforcement learning methods

Contents

This chapter tackles the first step towards answering the research question posited by this thesis, namely determining whether one can assess the efficacy of algorithms that coordinate residential energy flexibility. An environment is presented to test

MARL algorithms for the coordination of residential energy flexibility. A generative adversarial network (GAN)-based generator provides realistic EV consumption and at-home availability, PV generation, and household loads for ML applications. The environment[1] and data generator[2] are freely available on GitHub and will be used to perform experiments in chapters Chapters 4 to 6. The environment is modular, with different modelling modes available and possible user model customisation, input data changes and parameter adjustments. Any RL algorithm can then interact with this environment for benchmarking.

This chapter first introduces the convex optimisation formulation of the local energy system environment model in Section 3.1. Section 3.2 then translates this model into an RL environment MDP. Next, the Home Energy Data Generator (HEDGE) that will be used to generate input data is introduced in Section 3.3. Finally, other non-household data sources are presented in Section 3.4.

3.1 Local energy system convex optimisation model

This section describes the variables, objective function and constraints of the problem. This sets the frame for the application of the RL algorithms presented in Section 3.2.

Although this thesis focusses on RL, the convex optimisation is part of the environment as (a) a validation tool for the environment, (b) to provide an upper bound to the performance of the RL agents, and (c) to provide training datasets to agents.

3.1.1 Variables

Let us consider a set of time steps $t \in \mathcal{T} = \{t_0, ..., t_{\text{end}}\}$, a set of homes $i \in \mathcal{N} = \{1, ..., n\}$, a set of buses[3] $j \in \mathcal{B}$, and of lines (j, k) in $\mathcal{L}$. Decision variables are *italicised* and input data are written in roman. Energy units are used unless specified otherwise. Participants have an EV with vehicle-to-grid (V2G) capabilities, a PV panel, electric space heating and generic flexible household loads.

[1] `https://github.com/floracharbo/MARL_local_electricity`
[2] `https://github.com/floracharbo/hedge`
[3] Homes i are located on buses j, but not all buses contain homes.

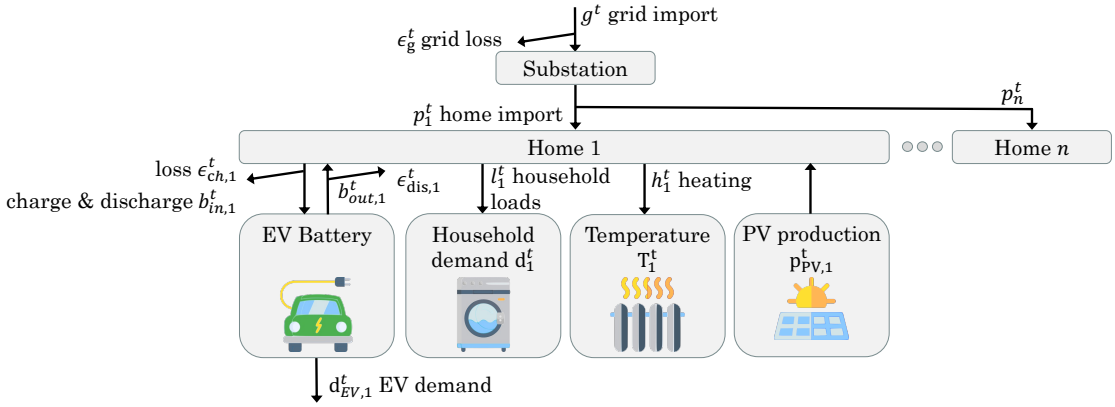


Figure 3.1: Local energy system model. Homes have vehicle-to-grid (V2G)-enabled electric vehicles, flexible household demand, electric heating, and PV generation. Energy balances apply at the asset-, home- and substation levels.

The EV at-home availability μ_i^t (1 if available, 0 otherwise), EV demand for required trips $d_{\mathrm{EV},i}^t$, household electric demand d_i^t, PV production $p_{\mathrm{PV},i}^t$, external temperature T_e^t, and solar heat flow rate Φ^t are specified as inputs for $t \in \mathcal{T}$ and $i \in \mathcal{N}$.

The local decisions by prosumers are the energy flows into and out of the battery $b_{\mathrm{in},i}^t$ and $b_{\mathrm{out},i}^t$, the reactive power provided by the EV battery $q_{\mathrm{EV},i}^t$, the electric heating consumption h_i^t and the household electric loads l_i^t. As illustrated in Figure 3.1, these have both local and system impacts. Local impacts include battery energy levels E_i^t, losses $\epsilon_{\mathrm{ch},i}^t$ and $\epsilon_{\mathrm{dis},i}^t$, prosumer import p_i^t, building mass temperature $T_{\mathrm{m},i}^t$ and indoor air temperature $T_{\mathrm{air},i}^t$. System impacts arise through the costs of total grid import g^t, distribution network trading and voltage levels v_j.

3.1.2 Objective function

The systems costs that prosumers cooperate to minimise can include grid (c_g^t), distribution (c_d^t), battery (c_b^t), substation (c_s^t) and voltage (c_v^t) costs. This objective function will be maximised both in convex optimisations off-line – to provide an upper bound for the achievable objective function, and in some cases to provide information to the learners during the simulated learning phase – and in the learning of MARL policies for decentralised online implementation. The objective function is:

$$\max F = \sum_{\forall t \in \mathcal{T}} \hat{F}_t = \sum_{\forall t \in \mathcal{T}} - \left(c_{\mathrm{g}}^t + c_{\mathrm{d}}^t + c_{\mathrm{b}}^t + c_{\mathrm{v}}^t \right) \tag{3.1}$$

Grid energy costs are defined as:

$$c_{\mathrm{g}}^t = \mathrm{C}_{\mathrm{g}}^t \left(g^t + \epsilon_{\mathrm{g}}^t \right) \tag{3.2}$$

Where losses incurred by imports and exports from and to the main grid are approximated as:

$$\epsilon_{\mathrm{g}}^t = \frac{\mathrm{R}}{\mathrm{V}^2} \left(g^t \right)^2 \tag{3.3}$$

And the grid cost coefficient $\mathrm{C}_{\mathrm{g}}^t$ is the sum of the grid electricity price and the product of the carbon intensity of the generation mix at time t and the Social Cost of Carbon, which reflects the long-term societal cost of emitting greenhouse gases [110]. The impacts of local decisions on upstream energy prices are neglected. Grid losses are approximated using the nominal root mean square grid voltage V and the average resistance between the main grid and the distribution network R [138], based on the assumption of small network voltage drops and relatively low reactive power flows [235]. The second-order dependency disincentivises large power imports and exports, which helps ensure interactions of transmission and distribution networks do not reduce system stability.

Distribution costs c_{d}^t are proportional to the distribution charge C_{d} on exports.

$$c_{\mathrm{d}}^t = \mathrm{C}_{\mathrm{d}} \sum_{i \in \mathcal{N} : p_i^t < 0} - p_i^t \tag{3.4}$$

where p_i^t corresponds to home-level imports if positive, home-level exports if negative. The resulting price spread between individual imports and exports decreases network constraints violation risks by incentivising the use of local flexibility first [67].

EV battery depreciation costs c_{b}^t are assumed to be proportional to throughput using the depreciation coefficient C_{b}, assuming a uniform energy throughput degradation rate [236].

$$c_{\mathrm{b}}^t = \mathrm{C}_{\mathrm{b}} \sum_{i \in \mathcal{N}} \left(b_{\mathrm{in},i}^t + b_{\mathrm{out},i}^t \right) \tag{3.5}$$

Voltage costs c_{v}^t are proportional to the voltage cost coefficients $\mathrm{C}_{\overline{\mathrm{v}}}$ and $\mathrm{C}_{\underline{\mathrm{v}}}$ which penalise voltage values above and below the maximum and minimum voltage levels $\overline{\mathrm{v}}$ and $\underline{\mathrm{v}}$ at each bus, respectively.

$$c_{\mathrm{v}} = \sum_{j \in \mathcal{B}:\ v_j^t > \overline{\mathrm{v}}} \mathrm{C}_{\overline{\mathrm{v}}} \left(v_j^t - \overline{v} \right) + \sum_{j \in \mathcal{B}: v_j^t < \underline{\mathrm{v}}} \mathrm{C}_{\underline{\mathrm{v}}} \left(\underline{v} - v_j^t \right) \tag{3.6}$$

3.1.3 Home-level constraints

Constraints for steps $\forall\ t \in \mathcal{T}$ and homes $\forall\ i \in \mathcal{N}$ are:

- Home-level energy balance – for η_{ch} and η_{dis} the charge and discharge efficiencies:

$$p_i^t = l_i^t + h_i^t + \frac{b_{\mathrm{in},i}^t}{\eta_{\mathrm{ch}}} - \eta_{\mathrm{dis}} b_{\mathrm{out},i}^t - \mathrm{p}_{\mathrm{PV},i}^t \tag{3.7}$$

- Battery energy balance:

$$E_i^{t+1} = E_i^t + b_{\mathrm{in},i}^t - b_{\mathrm{out},i}^t - \mathrm{d}_{\mathrm{EV},i}^t \tag{3.8}$$

- Battery charge and discharge constraints – for E_0, $\underline{\mathrm{E}}$ and $\overline{\mathrm{E}}$ the initial, minimum and maximum battery energy levels, and $\overline{\mathrm{b_{in}}}$ and $\overline{\mathrm{b_{out}}}$ the maximum charge and discharge per time step:

$$\mathrm{E}_0 = E_i^{t_0} = E_i^{t_{\mathrm{end}}} + b_{\mathrm{in},i}^{t_{\mathrm{end}}} - b_{\mathrm{out},i}^{t_{\mathrm{end}}} - \mathrm{d}_{\mathrm{EV},i}^{t_{\mathrm{end}}} \tag{3.9}$$

$$\mu_i^t \underline{\mathrm{E}}_i \leq E_i^t \leq \overline{\mathrm{E}}_i \tag{3.10}$$

$$b_{\mathrm{in},i}^t \leq \mu_i^t \overline{\mathrm{b_{in}}} \tag{3.11}$$

$$b_{\mathrm{out},i}^t \leq \mu_i^t \overline{\mathrm{b_{out}}} \tag{3.12}$$

- Consumption flexibility – the demand $\mathrm{d}_{i,m}^{t_{\mathrm{D}}}$ is met by the sum of partial household load consumptions $\hat{l}_{i,m,t_{\mathrm{C}},t_{\mathrm{D}}}$ at time t_{C} by prosumer i for load of type m (fixed or flexible) demanded at t_{D}. The flexibility boolean $\lambda_{i,m,t_{\mathrm{C}},t_{\mathrm{D}}}$ indicates if time t_{C} lies within the acceptable range $\{t_{\mathrm{D}}, ..., t_{\mathrm{D}+\mathrm{n_{flex}}}\}$ to meet $\mathrm{d}_{i,m}^{t_{\mathrm{D}}}$.

$$\sum_{t_{\mathrm{C}} \in \mathcal{T}} \hat{l}_{i,m,t_{\mathrm{C}},t_{\mathrm{D}}} \lambda_{i,m,t_{\mathrm{C}},t_{\mathrm{D}}} = \mathrm{d}_{i,m}^{t_{\mathrm{D}}} \tag{3.13}$$

- Consumption – the total consumption at time t_{C} is the sum of all partial consumptions $\hat{l}_{i,m,t_{\mathrm{C}},t_{\mathrm{D}}}$:

$$\sum_{t_{\mathrm{D}} \in \mathcal{N}} \hat{l}_{i,m,t_{\mathrm{C}},t_{\mathrm{D}}} = c_{i,m}^{t_{\mathrm{C}}} \tag{3.14}$$

- Heating – a Crank-Nicholson scheme [237] was reformulated so that the building mass and air temperature are expressed as a linear recursive expression of the previous building mass temperature, external temperature and heating energy. Here, ξ a 2x5 coefficients matrix, and $\underline{\mathrm{T}}_i^t$ and $\overline{\mathrm{T}}_i^t$ are the lower and upper temperature bounds. The workings to obtain this equation are included in Appendix B.

$$\begin{bmatrix} T_{\mathrm{m},i}^{t+1} \\ T_{\mathrm{air},i}^{t+1} \end{bmatrix} = \xi \left[1, T_{\mathrm{m},i}^t, \mathrm{T}_{\mathrm{e}}^t, \Phi^t, h_i^t\right]^\intercal \tag{3.15}$$

$$\underline{\mathrm{T}}_i^t \leq T_{\mathrm{air},i}^t \leq \overline{\mathrm{T}}_i^t \tag{3.16}$$

- Non-negativity constraints:

$$l_i^t, h_i^t, E_i^t, b_{\mathrm{in},i}^t, b_{\mathrm{out},i}^t, \hat{l}_{i,l,t_{\mathrm{C}},t_{\mathrm{D}}} \geq 0 \tag{3.17}$$

While the proposed framework could accommodate the use of idiosyncratic satisfaction functions to perform trade-offs between flexibility use and users' comfort, no such trade-offs are considered in this thesis, with comfort requirements for temperature and EV usage always being met. Field evaluations have shown that programmes that do not maintain thermal comfort are consistently overridden, increasing overall energy use and costs [58], while interference in consumption patterns and temperature set-points cause dissatisfaction [9]. Meeting fixed domestic loads, ensuring sufficient charge for EV trips, and maintaining comfortable temperatures are therefore set constraints.

3.1.4 Network constraints

There are two environment modes for network constraints modelling: substation energy balance, and power flow modelling.

Substation energy balance In the standard mode, the substation-level energy balance is modelled as:

$$\sum_{i \in \mathcal{N}} p_i^t = g^t \tag{3.18}$$

This is provided as an approximation in order to test optimisation-informed learning in a scalable manner, as the power flow modelling representation below is more computationally expensive. This approximation is used to compare the learning algorithms' coordination performance and develop the appropriate methodologies in Chapters 4 and 5, before the power flow modelling representation is used to evaluate impacts in more detail in Chapter 6. While this mode ignores power losses and voltage constraints, the coordination results of the algorithms developed under the substation-level energy balance hold when tested with the power flow modelling mode.

Power flow modelling Alternatively, the proposed testing and benchmarking environment also has a power flow modelling mode to replace Equation (3.18). This mode includes real and reactive power flow, losses and voltage modelling and will be used in Chapter 6.

The optimal power flow (OPF) problem is generally non-convex. In order to improve computational feasibility and tractability, a second-order cone relaxation [238] of the branch flow model [239] is adopted. This relaxed power relaxation is exact for radial networks, under some simple conditions that hold in many real distribution systems [240].

For line resistances $\mathrm{\rho}_{jk}$ and reactances x_{jk}, squared current I_{jk}, real and reactive net loads p_j^t and q_j^t, and real and reactive power flows P_{jk}^t and Ω_{jk}^t, the branch flow model is formulated as follows:

$$p_j^t = P_{jk}^t - \mathrm{\rho}_{jk} l_{jk}^t - \sum_{m:(k,m)\in\mathcal{L}} P_{km}^t, \; k \in \mathcal{B} \tag{3.19}$$

$$q_j^t = \Omega_{jk}^t - \mathrm{x}_{jk} l_{jk}^t - \sum_{m:(k,m)\in\mathcal{L}} \Omega_{km}^t, \; k \in \mathcal{B} \tag{3.20}$$

$$w_k^t = w_j^t - 2\left(\mathrm{\rho}_{jk}P_{jk}^t + \mathrm{x}_{jk}\Omega_{jk}^t\right) + \left(\mathrm{\rho}_{jk}^2 + \mathrm{x}_{jk}^2\right) l_{jk}^t,\ (j,k) \in \mathcal{L} \tag{3.21}$$

Where $w_j^t = |V_j^t|^2$.

The exact equation for the squared current is:

$$I_{jk}^t = \frac{(P_{jk}^t)^2 + (\Omega_{jk}^t)^2}{w_j^t},\ (j,k) \in \mathcal{L} \tag{3.22}$$

This quadratic equality constraint makes this OPF problem NP-hard [240]. This constraint can therefore be relaxed into the inequality:

$$I_{jk}^t \geq \frac{(P_{jk}^t)^2 + (\Omega_{jk}^t)^2}{w_j^t},\ (j,k) \in \mathcal{L} \tag{3.23}$$

The line losses can then be computed as $\mathrm{\rho}_{jk} l_{j,k}^t$.

This relaxed OPF (ROPF) provides a lower bound on OPF but is exact under conditions detailed in [240], which hold in in the case studies in this thesis.

While a standard OPF formulation minimises losses and generation costs, the objective function in the problem described in this thesis includes not only losses but also other terms described in Section 3.1.2. Moreover, the network constraints are coupled with the other decision variables in the problem, as detailed below.

Firstly, although the OPF Equations (3.19) to (3.23) above refer to buses $j \in \mathcal{B}$, there may be more buses than flexible homes modelled in Section 3.1.3, depending on the experiment. The total vector of buses' net loads $\boldsymbol{p}_{\text{all}}^t = \{p_j^t\ \forall\ j \in \mathcal{B}\}$ therefore contains the real power imports/exports at the home level both for flexible homes (decision variables) $\boldsymbol{p}_{\text{flex}}^t = \{p_i^t\ \forall\ i \in \mathcal{N}\}$ and fixed loads $\mathbf{p}_{\text{fixed}}^t = \{\mathrm{p}_{\text{fixed},i}^t\ \forall\ i \in \mathcal{N}'\}$, where $\mathcal{N}'$ is the set of non-flexible loads on the network and $\mathcal{B} = \mathcal{N} \cup \mathcal{N}'$.

For $\mathrm{K}_{\text{fixed}}$ a $[\mathrm{n}_{\text{buses}} \times \mathrm{n}_{\text{fixed buses}}]$ matrix of connection of fixed buses on the network, and K_{flex} a $[\mathrm{n}_{\text{buses}} \times \mathrm{n}_{\text{flex buses}}]$ matrix of connection of flexible buses on the network:

$$\boldsymbol{p}_{\text{all}}^t = \mathrm{K}_{\text{fixed}}\boldsymbol{p}_{\text{fixed}}^t + \mathrm{K}_{\text{flex}}\boldsymbol{p}_{\text{flex}}^t \tag{3.24}$$

Similarly, for the reactive powers, $\boldsymbol{q}_{\text{all}}^t = \{q_j^t\ \forall\ j \in \mathcal{B}\}$, $\boldsymbol{q}_{\text{flex}}^t = \{q_i^t\ \forall\ i \in \mathcal{N}\}$ and $\mathbf{q}_{\text{fixed}}^t = \{\mathrm{q}_{\text{fixed},i}^t\ \forall\ i \in \mathcal{N}'\}$, and:

$$\boldsymbol{q}_{\text{all}}^t = \mathrm{K}_{\text{fixed}}\boldsymbol{q}_{\text{fixed}}^t + \mathrm{K}_{\text{flex}}\boldsymbol{q}_{\text{flex}}^t \tag{3.25}$$

Moreover, the boundary condition of the power flow is linked to the total grid import as:

$$\sum_{\forall j \in \mathcal{B}:(0,j)\in\mathcal{L}} P_{0j}^{t} = g^{t} \tag{3.26}$$

Finally, even with the relaxation, the high computational complexity of the off-line non-convex optimisations can become a burden at scale, especially when the optimisations are necessary for constructing training datasets.

We therefore apply the Corrected DistFlow (CoDistFlow) approach [241] that handles the non-convexity of the original OPF problem by linearising the DistFlow power flow equations via replacing their non-linear terms with constants denoted as correction terms. It then iteratively improves the approximation by updating the values of the correction terms for the line losses and node voltages. The algorithm is iterated until the correction terms and the voltage magnitudes converge.

A limitation of this model is that all homes are assumed to be on a single phase, and the statistical approach to control adopted in this thesis does not include phase imbalance control.

3.1.5 Reactive power provision

In this section, we assume that the power factor of the heating, household loads, and PV generation is a fixed input. The EV battery can however be controlled by the RL policies if the "power factor control" mode is activated (and the power flows are modelled).

The reactive power imported and exported at the home level is defined as:

$$q_{i}^{t} = q_{\mathrm{EV},i}^{t} + \left(l_{i}^{t} + h_{i}^{t} - \mathrm{p}_{\mathrm{PV},i}^{t}\right) \tan(\cos^{-1}(PF)) \tag{3.27}$$

Fixed power factor For a fixed power factor PF, the reactive power imported and exported by the EV battery is computed as:

$$q_{\mathrm{EV},i}^{t} = p_{\mathrm{EV},i}^{t} \tan(\cos^{-1}(PF)) \tag{3.28}$$

Where the real power EV battery output $p_{\mathrm{EV},i}^{t}$ is:

$$p_{\mathrm{EV},i}^{t} = \frac{b_{\mathrm{in},i}^{t}}{\eta_{\mathrm{ch}}} - b_{\mathrm{out},i}^{t}\eta_{\mathrm{dis}} \tag{3.29}$$

Power factor control Alternatively, a battery power factor control mode can be activated to replace Equation (3.28). When the power factor control is enabled, the reactive power supplied by the battery $q^t_{\text{EV},i}$ is controlled within the constraints of the maximum apparent power capacity of the EV battery S_{EV} and by the EV at-home availability μ^t_i.

$$(q^t_{\text{EV},i})^2 + (p^t_{\text{EV},i})^2 \leq \overline{S_{\text{EV}}} \tag{3.30}$$

The reactive power can only be provided if the EV is available at home:

$$q^t_{\text{EV},i} \leq \mu^t_i M \tag{3.31}$$

where M is a large number.

3.2 Reinforcement learning environment representation

This section now presents the MDP RL environment representation, which directly mirrors the convex optimisation formulation defined in Section 3.1. In the MARL approach, independent prosumers learn to make individual decisions, which together aim to maximise the statistical expectation of the objective function in Section 3.1.2. Actions are implemented one time step at a time, in contrast to the "omniscient", day-ahead global optimisation representation in Section 3.1.

At time step $t \in \mathcal{T}$, each agent is in a state $s^t_i \in \mathcal{S}$ corresponding to accessible observations in the environment and selects an action $a^t_i \in \mathcal{A}$ as defined in Section 3.2.2. This action dictates the decision variables in Section 3.1.1 $b^t_{\text{in},i}$, $b^t_{\text{out},i}$, h^t_i and l^t_i. The environment then produces a reward $r^t \in \mathcal{R}$ and agents transition to a state s^{t+1}_i. Agents learn individual policies π_i by interacting with the environment.

3.2.1 States

The potential states accessible to agents to inform action selection in the testing and benchmarking environment are tabulated in Table 3.1. The grid cost coefficients (over present and future time steps) are indentified as the most valuable state to

inform the RL agents' decision-making in Sections 4.3.2 and 5.3.1 and will be used as the unique state in most experiments in this thesis. The lowest voltage value was also shown to be a valuable observation for cooperative agents to manage voltage deviations in real time (Section 6.1.2).

Table 3.1: Environment states.

State	Definition
Grid cost coefficient $\mathrm{C}_{\mathrm{g}}^{t}$	The sum of the grid electricity price and the product of the carbon intensity of the generation mix at time t and the social cost of carbon.
Grid cost coefficient in the next n time steps	$\{\mathrm{C}_{\mathrm{g}}^{t}, ..., \mathrm{C}_{\mathrm{g}}^{t+n-1}\}$
Time	Time step within the current day
E_i^t	Battery level at the start of the time step
μ_i^t	EV at-home availability for the current time step
$\mathrm{d}_{\mathrm{EV}}^{t}$	EV demand for the current time step
τ_{EV}	$= \frac{\text{increase in charge level required for next trip}}{\text{number of time steps until next trip}}$
Day type	Weekday or weekend day
Flexibility	Amount of flexibility available at the current time step, in energy units, between the minimum and maximum possible home import/export given local temperature, flexible loads and battery constraints $F_i^t = \overline{p_i^t} - \underline{p_i^t} = (\overline{\Delta s_i^t} - \underline{\Delta s_i^t}) + (\overline{h_i^t} - \underline{h_i^t}) + (\overline{l_i^t} - \underline{l_i^t})$ where $(\overline{p_i^t}, \underline{p_i^t})$, $(\underline{\Delta s_i^t}, \overline{\Delta s_i^t})$, $(\underline{h_i^t}, \overline{h_i^t})$ and $(\underline{l_i^t}, \overline{l_i^t})$ are the lower and upper bound for respectively the home imports, change in EV battery level, heating energy consumption, and household consumption for home i and time t.
Flexibility boolean	Whether there is any flexibility between the minimum and maximum possible home import/export at the current time step given local constraints
Storage flexibility	Amount of flexibility available between the minimum and maximum battery charge at the current time step.
Storage flexibility boolean	Whether there is any flexibility at the current time step between the minimum and maximum battery charge.
Lowest voltage	Minimum voltage value across all network buses at the start of each time step. This is an indicator of the accumulation of under-voltages in the distribution network.

3.2.2 Actions

Individual actions There are three active power RL agent actions, as well as an optional EV battery power factor action. The constraints are then enforced in the environment thanks to a physics-informed approach to translating these agent actions into feasible EV, heating and flexible consumption variables. This method of "clipping" an action after it is selected by the agent improves computational efficiency [137].

- Battery action $a^t_{\mathrm{EV},i} \in [-1, 1]$: for negative values scales the maximum amount of feasible discharge ($b^t_{\mathrm{out},i}$), and for positive values of feasible charge ($b^t_{\mathrm{in},i}$), as illustrated in Figure 3.2. Given Equations (3.8) to (3.12):

$$(\underline{\Delta s^t_i}, \overline{\Delta s^t_i}) = f(E^t_i, d^t_{\mathrm{EV},i}, ..., d^{t_{\mathrm{end}}}_{\mathrm{EV},i}, \mu^t_i, ..., \mu^{t_{\mathrm{end}}}_i, \mathrm{E}_0, \underline{\mathrm{E}}, \overline{\mathrm{E}}, \overline{\mathrm{b}_{\mathrm{out}}}, \overline{\mathrm{b}_{\mathrm{in}}}) \tag{3.32}$$

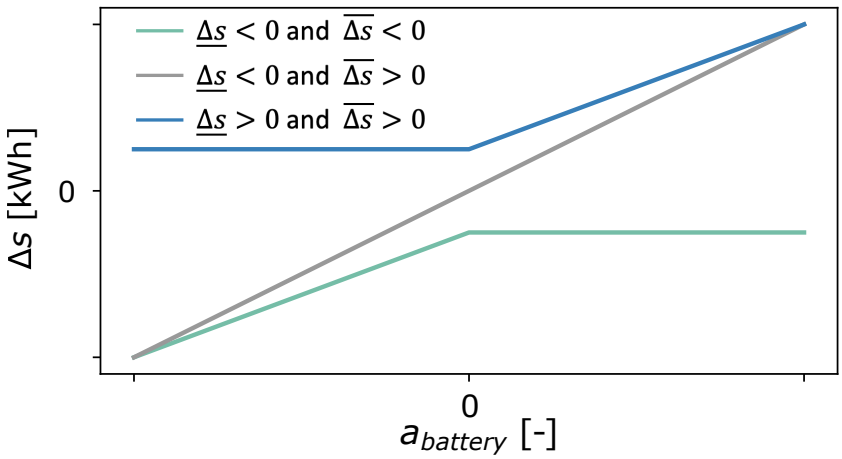


Figure 3.2: Translating the battery action into the change in battery energy level.

- Heating action $a^t_{\mathrm{heat},i} \in [0, 1]$: linearly scales between the minimum and maximum amount of feasible heating (h^t_i) given local temperature constraints. Given Equations (3.15) and (3.16):

$$(\underline{h^t_i}, \overline{h^t_i}) = f(T^t_{\mathrm{m},i}, \mathrm{T}^t_{\mathrm{e}}, \Phi^t, \overline{T^t_{i,\mathrm{air}}}, \underline{T^t_{i,\mathrm{air}}}) \tag{3.33}$$

$$h^t_i = (1 - a^t_{i,\mathrm{heat}})\underline{h^t_i} + a^t_{i,\mathrm{heat}}\overline{h^t_i} \tag{3.34}$$

- Household consumption action $a^t_{\text{cons},i} \in [0, 1]$: scales feasible flexible household consumption (l^t_i). Given Equations (3.13) and (3.14):

$$(\underline{l^t_i}, \overline{l^t_i}) = f(d^{t-\text{n}_{\text{flex}}}_i, ..., d^t_i, c^{t-\text{n}_{\text{flex}}}_i, ..., c^{t-1}_i) \tag{3.35}$$

$$l^t_i = (1 - a^t_{i,\text{cons}})\underline{l^t_i} + a^t_{i,\text{cons}}\overline{l^t_i} \tag{3.36}$$

- Reactive power action $a^t_{\text{reactive},i} \in [-1, 1]$: scales the maximum amount of reactive power injection or consumption $\overline{q_{\text{EV}}}^t_i$ given by Equation (3.30).

Note that although these actions are continuous, they can be discretised into intervals for tabular implementations.

Combined action Alternatively, a unique combined action representation was developed in the environment. While this did not improve performance in this thesis (see Figure 4.1), it may be useful for learning methodologies that perform best with small action spaces.

This formulation is presented in Appendix C.

3.2.3 Reward

The global reward $r^t \in \mathcal{R}$ for each time step t corresponds to the share $\hat{F}_t$ of the system objective function presented in Section 3.1.2.

3.2.4 Step function

The computational requirements of executing actions chosen by the RL agent are minimal. Once the training is over, the home must only receive a price signal, perform a feed-forward using the previously trained policy weights (if experimenting with neural networks – the number of flops required will depend on the neural network size) or look up the action in a Q-table. Thus, despite receiving solely an individual price signal, the home's objective is not to maximise its individual rewards based on this signal. Rather, it endeavours to maximise joint rewards by leveraging the knowledge of appropriate coordinated responses acquired through

cooperative training, thereby determining the appropriate actions to undertake in response to various states (price signals).

To convert the RL policy action into local decisions, the agent then requires information on their current PV generation, battery level, flexible loads and indoor air temperature, as described below in Section 3.2.2. Then, the step function simply:

- Evaluate the maximum and minimum heating energy at the current time step to stay within comfortable temperature bounds given internal and external temperatures
- Evaluate the maximum and minimum EV battery charge and discharge given the current and future EV demand and at-home availability
- Evaluate the maximum and minimum flexible household consumption
- Translate RL actions $\in [-1, 1]$ or $[0, 1]$ into the appropriate household consumption, EV charge and discharge, and heating control values.

This sequence can be efficiently computed at the domestic level using inexpensive chips with minimal energy demands, as it necessitates fewer than 1 kilo floating point operations (kFLOPs). This was a crucial prerequisite for the computational scalability of the coordination mechanisms developed in this thesis, as adherence to the third criterion outlined in Section 1.2 necessitates the imposition of a minimal and constant computational burden as system size increases.

The environment then updates its variables, such as the battery level, temperature and flexible loads. Finally, the network impacts of the control actions are directly simulated using pandapower, an open-source Python tool that offers a Newton-Raphson power flow solver formerly based on PyPower, which has been accelerated with just-in-time compilation [242].

During implementation, only local data is needed, with no sharing of personal data, behaviour and action choice. Global variables which are assumed to be independent of other agents' actions are also available, such as the electricity real-time price. The local data needs for implementation are:

- Electricity real-time prices
- Local battery level
- Local electricity demand, and share of electricity demand which is flexible
- Local PV generation
- Occupants' instructions: local vehicle travel schedule, heating temperature requirements

3.3 Home energy data generation tool

The Home Energy Data Generator (HEDGE) tool tackles the challenge of how to generate home energy consumption and generation data for use in data-driven algorithms. This open-access tool is trained on real data to generate realistic photovoltaic (PV) generation, household loads, and EV consumption and at-home availability profiles. The corresponding variables $p^t_{PV,i}$, d^t_i, $d^t_{EV,i}$, and μ^t_i are fed into the model presented in Section 3.1. Interaction with this data will shape the policies learned through RL [185] and should reflect resource intermittency and uncertainty to maximise the expectation of rewards in a robust way without over-fitting.

3.3.1 Objectives and motivation

The HEDGE tool tackles the challenge of how to generate home energy consumption and generation data for use in data-driven algorithms. This open-access tool can generate realistic PV generation, household loads, and EV consumption and at-home availability profiles.

The characterisation and simulation of residential energy resources are of increasing interest given their potential for demand-side response [243] described in Chapter 1. Chapter 2 has shown that data-driven methods are of particular interest in the field of distributed energy forecasting [244] and control [9] for three main reasons.

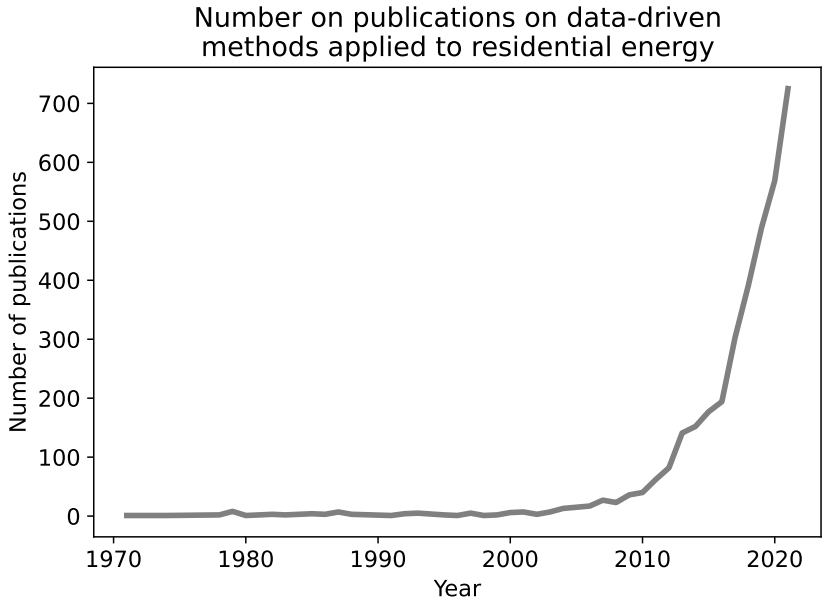


Figure 3.3: Number of publications on data-driven methods for residential energy.

Firstly, there is high uncertainty at the local level [245], due to the small scale of residential electricity consumption and generation, and their behavioural and weather dependencies. Secondly, there are limitations to personal data sharing, particularly in real time. This is due to both the limited availability of communication and computation infrastructure at the scale of an individual home and to the privacy requirements of the residential sector [246]. Thirdly, centralised optimisation methods have limited scalability [213, 247]. Therefore, data-driven analysis and control of the residential energy sector are of increasing interest [9, 213]. Figure 3.3 thus shows that the number of publications in the literature investigating the use of data-driven methods in the field of residential energy[4] has seen an exponential increase since 2000 (30% average yearly increase).

This section bridges the first research gap laid out in Section 2.3. A new tool is proposed that generates EV-, PV- and household demand-related data semi-

[4]Scopus keyword search selecting for publications whose title and abstract include at least one mention of each the residential sector (home or residential), of data-driven methods (data-driven, learning, big data, or forecasting), the energy sector (energy, electricity, power, voltage, renewable) and local energy appliances (car, PV, solar, loads, smart, IoT, storage, battery, heating, HVAC, generation, electric vehicle, EV, appliance, demand response, demand-side response, peer-to-peer, consumer, or fridge)

randomly based on large-scale real-life datasets, while preserving profile magnitude and behavioural consistency over time. While this model uses UK data, the model could be adapted to use similar data from other countries, so long as banks of data on household consumption, PV generation, and travel patterns are available.

Pre-training neural network weights for input data generation means researchers and practitioners do not need to download large databases (here, the raw databases that had to be downloaded were of size 40.12 GB) and run time- and computational resource-hungry data preparation and training steps. They only need to download the pre-trained weights (files of size 125 kB) and perform a feed-forward to obtain training and testing data.

The rest of this section is structured as illustrated in Figure 3.4. Section 3.3.2 presents the data generation mechanism. In Section 3.3.3, the mechanism used by HEDGE to generate data profiles is presented.

3.3.2 Data preparation

The data preparation steps are listed in Table 3.2 and detailed in the subsections below.

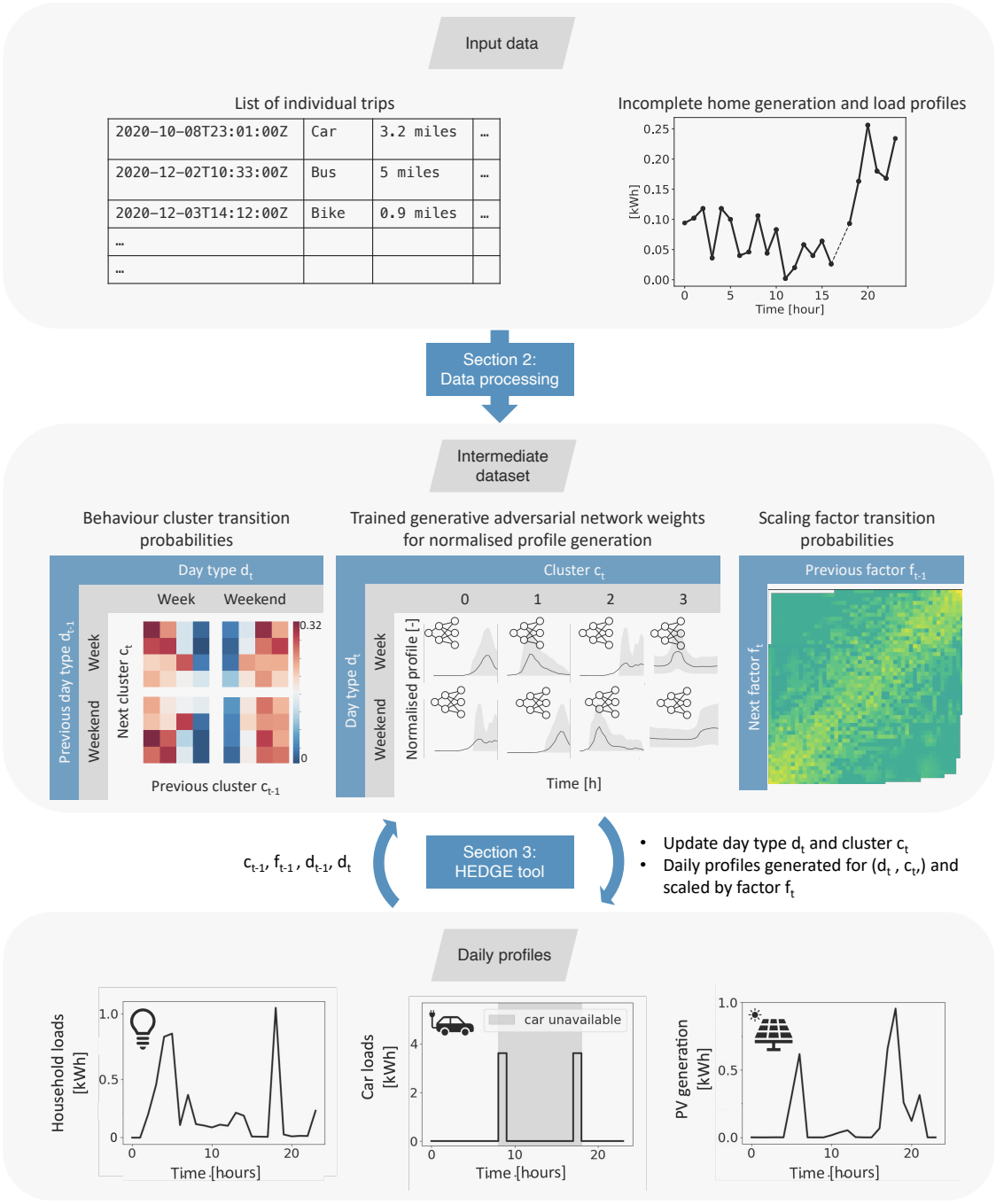


Figure 3.4: Workflow from raw input data to the generation of random realistic household energy profiles. The two main steps, 'Data processing' and 'HEDGE tool use', correspond to Sections 3.3.2 and 3.3.3.

	Solar generation	**Household loads**	**Electric vehicles**
1. Import data sources	Customer-led network revolution (CLNR) dataset TC1a [248]	CLNR dataset TC5 [249]	UK National Travel Survey [229]
2. Data selection and filtering	Only residential data is used, and for valid date ranges.		Only residential car journeys are selected.
3. Conversion to relevant daily profiles	Convert to resolution specified (which has to be greater than or equal to one minute)	Get resolution specified (greater than or equal to 30 minutes)	Convert lists of trips to distance travelled per time interval at the resolution specified. Infer the type of trip (motorway, urban, rural) from the location and distance. Convert the distance travelled to electricity consumption. Infer at-home availability.
4. Missing data interpolation	Linearly fill in single missing time steps or discard the day of data.		
5. Normalisation	Normalise daily profiles by the sum of the electricity consumption/generation over one day. Record the scaling factors.		
6. Behaviour grouping	No clustering – group by month	For each day type (weekday and weekend day), obtain 4 clusters using K-means.	For each day type (weekday and weekend day), obtain 3 clusters using K-means, as well as one for no-travel days.
7. Profile generation	Train generative adversarial networks (GANs) to generate realistic profiles for each behaviour group.		
8. Scaling factor transition characterisation	Using 50 discrete time intervals for each day type transition, obtain the discrete transition probabilities between subsequent days.		
9. Behaviour cluster transition characterisation	No clustering	Compute the transition probabilities for each cluster type and day type transition based on the real datasets.	

Table 3.2: Data preparation steps

3.3.2.1 Anonymised data selection and import

Anonymised load and PV generation profiles are obtained from the Customer-Led Network Revolution (CLNR), a UK-based smart grid demonstration project [248, 249], which collected data from 13,000 customers between 2011 and 2014. PV sources have nominal capacities between 1.35 and 2.02 kWp.

Anonymised mobility data from the National Travel Survey (NTS) [229] is used, from 105,912 Great Britain households between 2002 and 2020. The NTS surveys the general population's travel patterns and does not focus on EVs. This dataset was selected rather than EV trial data, as this offers a less biased view into the general population's travel patterns thanks to both the larger volume of data available, and because the self-selected EV early trial participants may not be representative of patterns once EVs become widely adopted. It is assumed that internal combustion engine (ICE) car travel patterns can be substituted for those of EVs, within battery constraints, as in [250].

To overcome memory issues and reduce computational time, the datasets are broken down into n segments, without interrupting data for single homes. Data size reduction steps such as data filtering and granularity adjustments are conducted first before merging the different streams.

A limitation of these datasets is that behaviour and load profiles may have evolved since the date of collection. For example, the use of incandescent rather than LED lights was more common historically, and work patterns have evolved. Moreover, the datasets were collected in the UK, and may not be representative of other countries. However, the methodology proposed could be used with other datasets for different contexts. Finally, the dataset does not provide information on the breakdown of the household loads. While the share of households using electric heating and possessing at-home EV chargers was low at the time of the data collection (2011-2014), it is possible that some heating and transport electrical loads may already be in the source data. There may therefore be a risk of double counting these loads if they are also modelled separately.

3.3.2.2 Data selection and filtering

Firstly, the measurements of interests are selected. In the case of the NTS data, only household car trips are conserved, and only homes classified as urban and rural are used. This is because the household type is needed to infer driving type and convert trips into electricity use at a later stage. Moreover, trips exceeding the maximum user-defined EV maximum hourly and daily energy demand are removed, as they would not be feasible with an electric car.

Then, the start and end times for data validity for each home are enforced and data beyond valid ranges discarded[5]. Data validity ranges are characterised by the start of valid time, the end of valid time, and the duration of valid time. If one of these is missing, it can be inferred. If two or more of these pieces of information are missing, the validity of the data cannot be confirmed, and it is discarded.

3.3.2.3 Conversion to relevant daily profiles

Sequences of subsequent data points for single homes are converted to the required resolution (e.g. hourly) and split into individual days.

In the case of CLNR data, this time granularity must be lower than that of the original data, e.g. one minute for PV generation and 30 minutes for household loads. Incomplete days with more than one consecutive data point missing are discarded.

In the case of the NTS travel data, lists of trips are converted to daily profiles of distance travelled. The at home-availability of the vehicles is then inferred from the recorded journeys' origin and destination. Equivalent EV energy consumption profiles are obtained using representative consumption factors from a tank-to-wheel model proposed in [250], dependent on travel speed and type (rural, urban, motorway). Motorway travel is assumed for trips larger than 10 miles [250].

[5]Enforcing a maximum of 50 kW at any given time step, and a daily maximum energy consumption of 200 kWh, 0.68% of the days of travel data considered were discarded

3.3.2.4 Missing data interpolation

For days with missing data points due to data recording or communication issues during the data collection, the options are either to interpolate the missing data points, or to discard the day of data entirely [251]. In this work, we discard days containing series of two or more subsequent data points missing, and we interpolate single missing data points. Continuous data profiles can be generated by filling in the missing data periods with imputed data [251]. This is so that we can increase the number of available full days of available data, making the HEDGE tool more representative of a wealth of real-life behaviours, while not compromising data quality.

To fill in single missing data points, the following options are tested:

- Linear interpolation between time steps before and after
- Select the point at the same time the day before or after (whichever has the lowest sum of squares of differences between the previous and subsequent point on the current day)
- Select the point at the same time one or two days before or after (whichever has the lowest sum of squares of differences between the previous and subsequent point on the current day)
- Select the point at the same time one day or week before or after (whichever has the lowest sum of squares of differences between the previous and subsequent point on the current day)

As shown in Figure 3.5, linearly interpolating results in the lowest average and 99th percentile. Therefore, this method is used to fill in missing data.

3.3.2.5 Normalisation

Normalisation is performed ahead of profile clustering and GAN training. Each daily profiles for energy generation and consumption are normalised such that $\sum_{t=0}^{24} x(t) = 1$, and the corresponding scaling factors are recorded. These profiles can then be

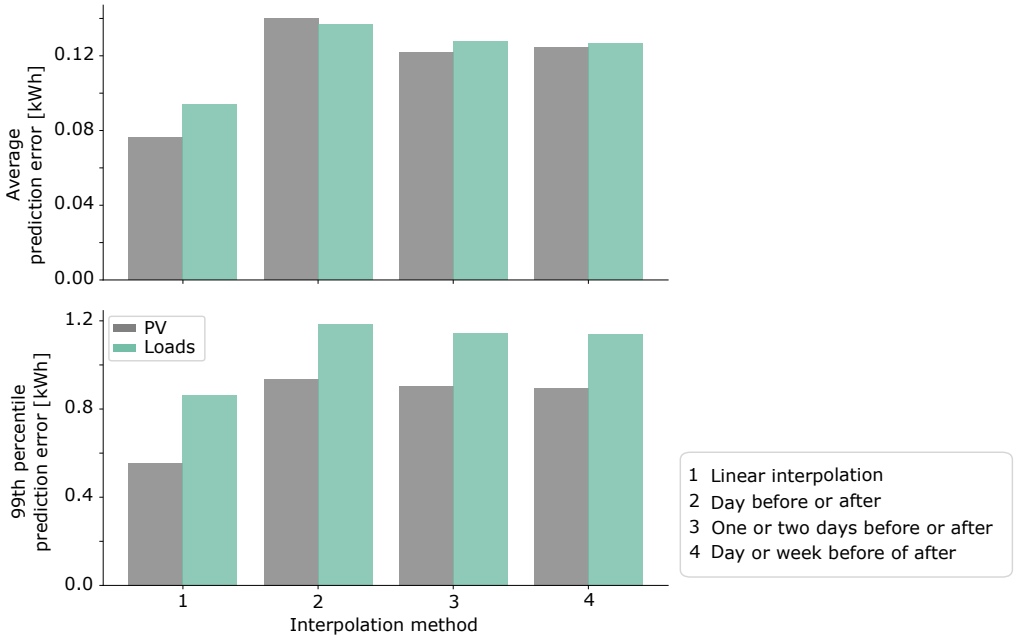


Figure 3.5: Comparison of interpolation methods.

scaled up consistently to match the expected total energy generation/consumption over a day for a given household by the generation tool, as further described in Sections Sections 3.3.2.9 and 3.3.3.

3.3.2.6 Behaviour clustering

For behaviour-dependent profiles, specifically household loads and EV patterns, the normalised profiles are grouped into clusters κ based on behavioural patterns for both weekday and weekend days. This clustering facilitates the creation of a repository of normalised profiles for each behaviour cluster group. This collection of profiles can subsequently serve as the foundation for training GANs to generate profiles representative of each cluster. When using HEDGE, different homes will have different likelihood of belonging to each behaviour group, and profiles can be generated accordingly to maintain consistency. Thus, the aim of generator is not to generate profiles corresponding to an "average" home, but rather for homes to represent different behaviour patterns present in the same proportion as in the original dataset.

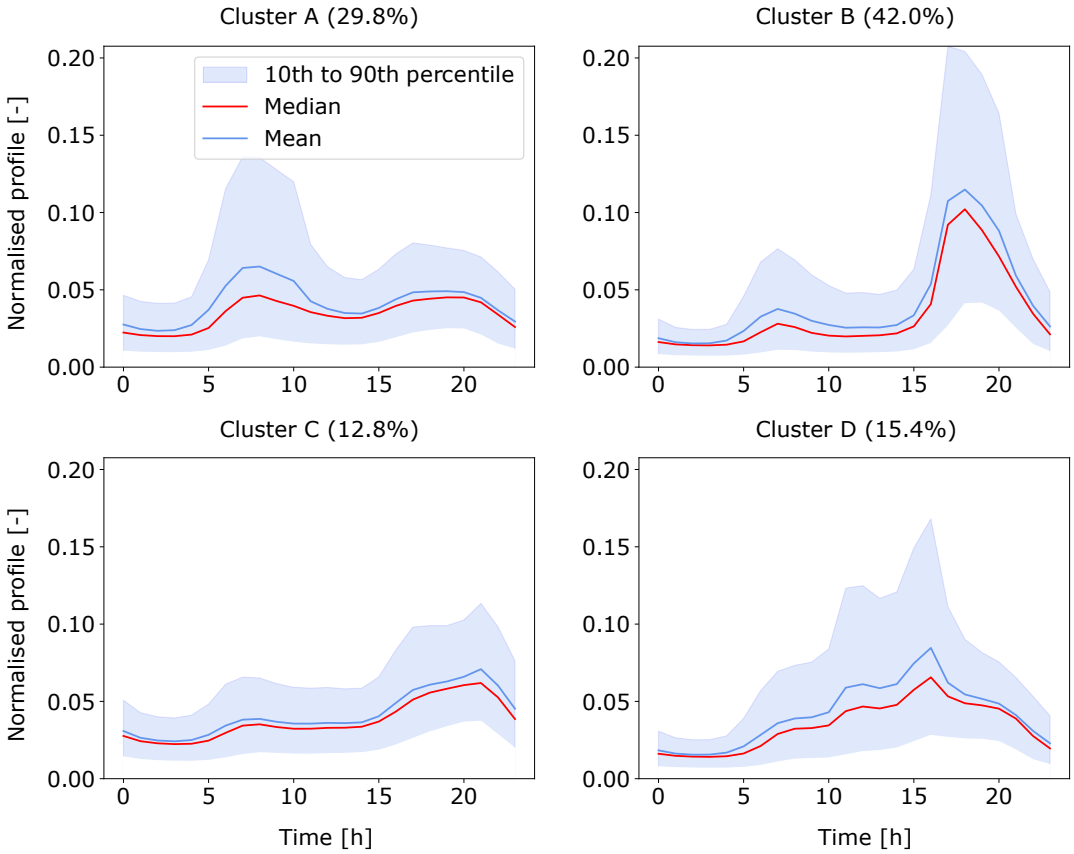


Figure 3.6: Four behaviour clusters for weekday household consumption normalised profiles.

K-means is used, minimising the within-cluster sum-of-squares [252] in four clusters κ for both weekday and weekend data (with one for no travel). The features used for load profiles clustering are normalised peak magnitude and time, and normalised values over critical time windows[6], and those for travel are normalised values between 6 am and 10 pm. PV profiles were grouped per month. The user can define the number of clusters as an input.

The weekday behaviour clusters for household loads and EV consumption are illustrated in Figures 3.6 and 3.7.

3.3.2.7 Profiles generation

Neural networks are then trained to generate populations of realistic normalised profiles corresponding to each behaviour cluster κ and day type ν. Pre-training

[6] 0-7 am, 7-11 am, 11 am-2 pm, 2-5 pm, 5-9 pm, 9-12 pm

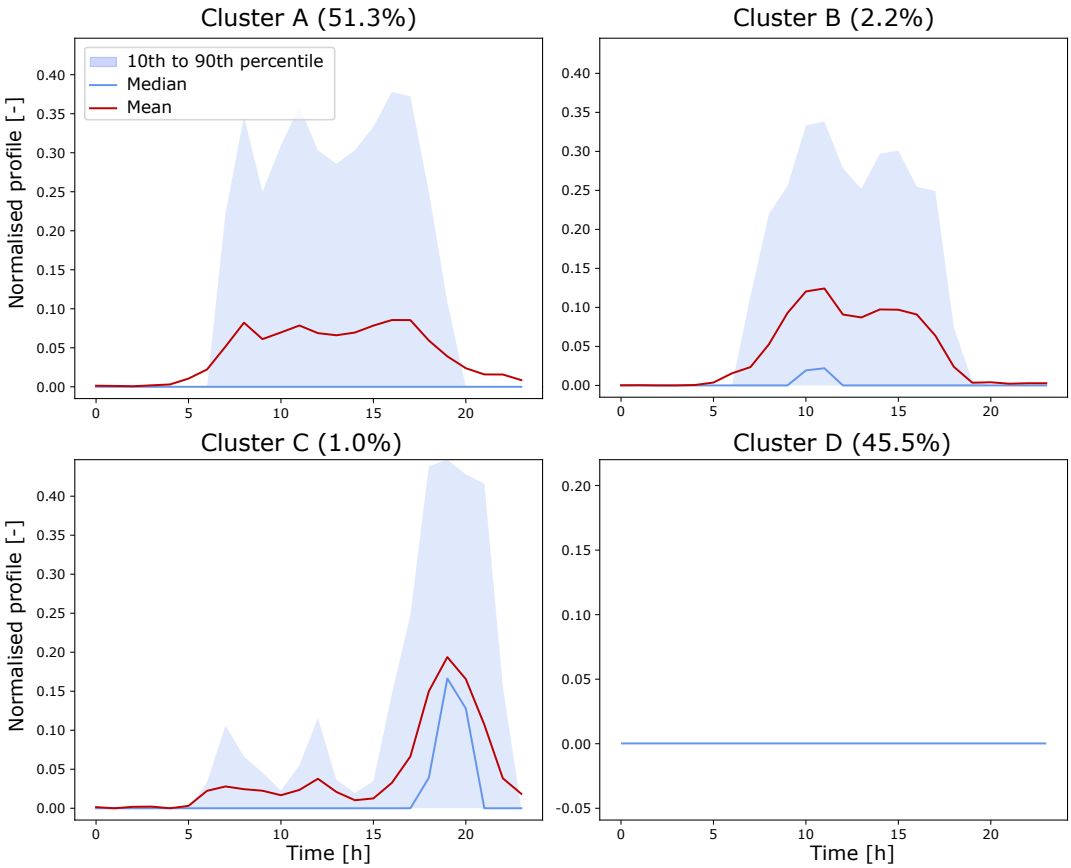


Figure 3.7: Four behaviour clusters for weekday EV consumption normalised profiles. The fourth cluster corresponds to days with no travel.

neural network weights means that researchers and practitioners do not need to download large databases (here, the raw databases that had to be downloaded were of size 40.12 GB) and run time- and computational resource-hungry data preparation and training steps. They only need to download the pre-trained weights (files of size 125 kB) and perform a feed-forward to generate realistic training and testing data using HEDGE.

As illustrated in Figure 3.8, GANs [253] consist of two simultaneously trained models. The generative model $\mathcal{G}$ takes as input a random noise vector z and produces fake data $x_{\text{synthetic}} = \mathcal{G}(z)$, aiming to fool the discriminator into thinking they are from the original dataset x_{real}. The discriminator model $\mathcal{D}$ takes as input data x and produces a probability score $\mathcal{D}(x) \in [0, 1]$ that indicates the likelihood that x is real data.

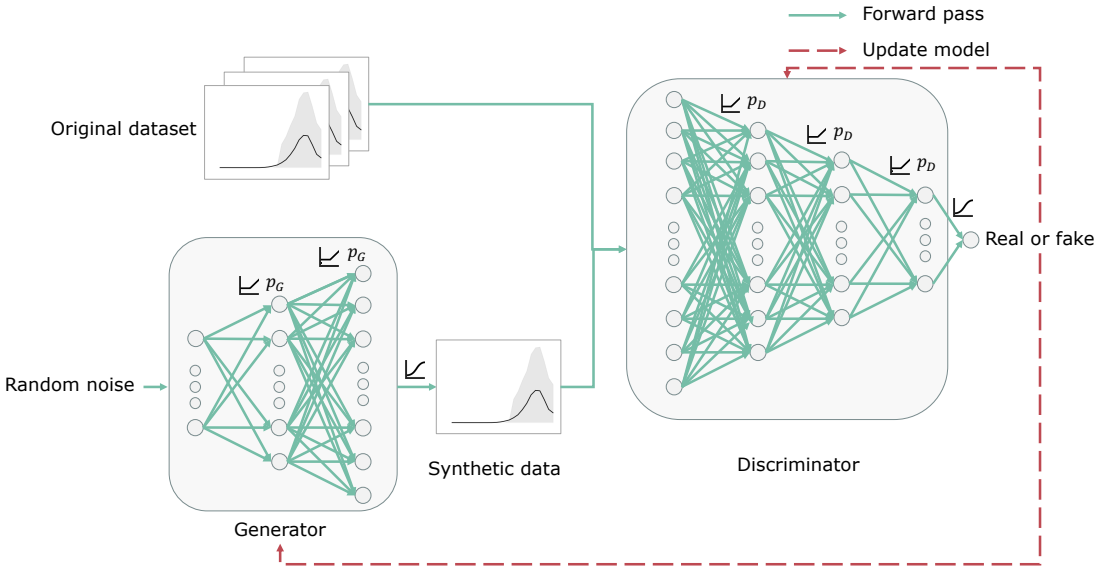


Figure 3.8: Generative adversarial networks architecture.

Each network aims to minimise the following losses during training:

- The discriminator aims to maximise the probability of correctly discriminating between the real real data and the fake data generated by the generator network $\mathcal{G}$, by minimising the binary cross-entropy between the real (1) and fake (0) labels and the probabilities assigned by the discriminator:

$$\ell_{\mathcal{D}} = -\mathbb{E}_{x_{\text{real}}} \left[\log \mathcal{D}(x_{\text{real}})\right] - \mathbb{E}_z \left[\log(1 - \mathcal{D}(\mathcal{G}(z)))\right] \tag{3.37}$$

- The generator loss is calculated from the discriminator's classification – it gets rewarded if it successfully fools the discriminator and gets penalised otherwise. The loss function aims to minimise the binary cross-entropy between the fake labels and the probabilities assigned to the fake generated data by the discriminator:

$$\ell_{\mathcal{G}} = -\mathbb{E}_z \left[\log(\mathcal{D}(\mathcal{G}(z)))\right] \tag{3.38}$$

To generate realistic populations of synthetic profiles, the following additional parameters and algorithm configurations are used:

- The learning rate is exponentially decayed to avoid oscillation and to obtain faster convergence [254], so that the learning rate at each epoch is:

$$\alpha_{\text{epoch}} = \alpha_0 \left(\frac{\alpha_{\text{end}}}{\alpha_0} \right)^{\frac{epoch}{n_{\text{epochs}}}} \tag{3.39}$$

- The positivity of the generator's output is enforced by using the sigmoid activation function [255] on the last layer of the generator network:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.40}$$

- Dropout layers [256] are employed within the neural network architectures to improve the performance of the models. This prevents overfitting to the training data by randomly dropping out (setting to zero) some of the outputs of the neurons during training, with probability $p_{\mathcal{G}}$ for the generator and $p_{\mathcal{D}}$ for the discriminator, effectively removing them from the network for that iteration. By doing this, the network becomes less sensitive to the specific weights of individual neurons and is forced to learn more robust features that are shared across multiple neurons.

Moreover, we further propose the following:

- A population of profiles $u \in \{1, \ldots, n_{\text{profiles}}\}$ is generated at each forward, pass, rather than one profile. This is to ensure that the GAN generates variability within one population that is realistic, rather than converging towards one realistic profile.

- A penalty is added to the generator's loss if the sum of the generated normalised profiles diverges from 1:

$$\ell_1 = W_1 \left(\frac{\sum_u \sum_t x_u^t}{n} - 1 \right)^2 \tag{3.41}$$

- A penalty is added to the generator's loss if the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles and the mean over the whole generated population for each time step t varies from the original dataset:

$$\ell_2 = W_2 \sum_{l \in \{10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, \text{mean}\}} \sum_t \left(x_l^t - x_{\text{real},l}^t \right)^2 \tag{3.42}$$

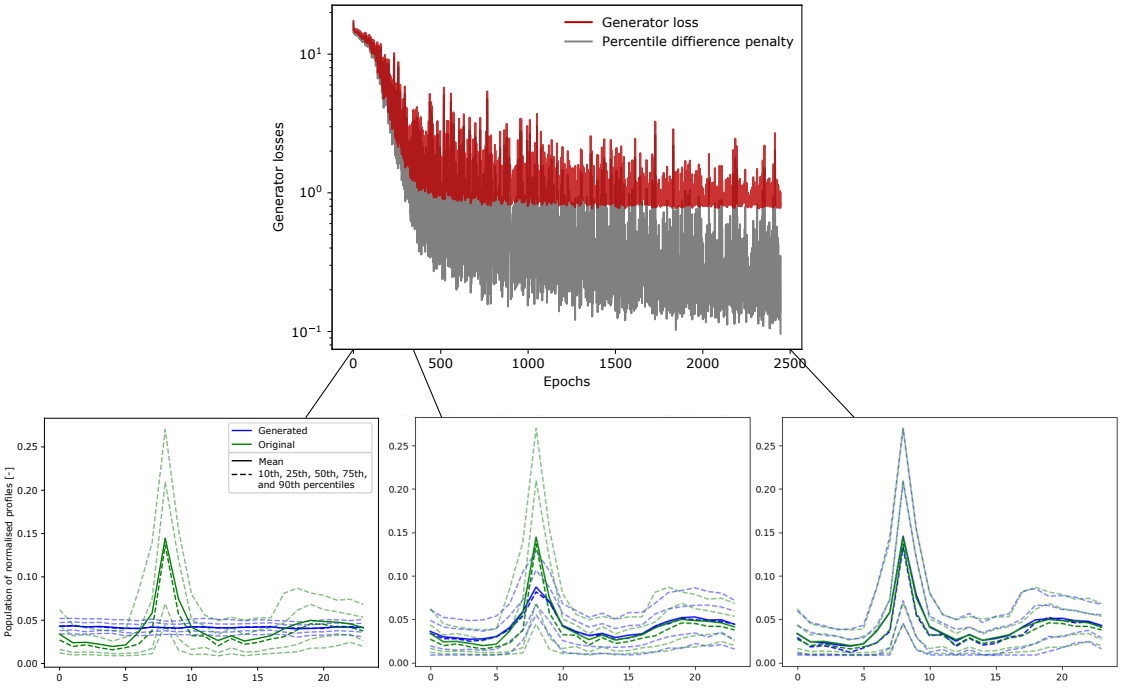


Figure 3.9: Example of generated populations of 50 household load normalised profiles against the distribution of the original dataset throughout the GAN training.

and

$$\ell'_G = \ell_G + \ell_1 + \ell_2 \tag{3.43}$$

- An exponentially decaying noise is added to the exploration, to improve the efficiency and effectiveness of learning by encouraging exploration, avoiding overfitting oscillation, and obtain faster convergence. The decay helps balance the exploration and exploitation trade-off over time. The noise at each epoch is thus:

$$\epsilon_{\text{epoch}} = \epsilon_0 \left(\frac{\epsilon_{\text{end}}}{\epsilon_0} \right)^{\frac{epoch}{n_{\text{epochs}}}} \tag{3.44}$$

The training parameters are tabulated in Table C.4.

3.3.2.8 Assessment of Generative Adversarial Networks

Assessing the performance of GANs can be challenging, especially for GANs generating time-series data, which is a more nascent field of study relative to

the computer vision domain. A combination of both qualitative and quantitative assessments is recommended [257].

Firstly, we therefore perform a qualitative visual assessment of the profiles generated by the GAN. An example of a generated population of 50 household load profiles throughout the training is presented in Figure 3.9. While the profiles generated before the training starts do not match the target distribution, the population of profiles that is generated at the end of the training visually matches the target population in terms both of mean and in terms of the distribution and variability of the population of profiles throughout the day. This shows that the generated profiles are diverse enough, as samples are distributed to cover the real data.

Secondly, we perform a quantitative evaluation, by adopting the "Train on Synthetic, Test on Real" (TSTR) framework proposed in [258] to evaluate the output of a GAN. This framework tests the usefulness of the GANs, by assessing the extent to which the generated data maintains the predictive attributes of the original. The testing sequence is as follows:

1. Split the real dataset into a training (80% of the data) and a testing (20%) dataset.

2. Train the GANs using the training dataset.

3. Generate synthetic data with the GANs.

4. Train a model using the synthetic data – Here, we train a classifier which aims at predicting which cluster a population of data profiles belongs to.

5. Test the classifier model using the held-out testing data. By determining the classifier's quality, this evaluation method, in turn, aims to assess the quality of the generated data in being used for real applications.

Similar to the TSTR method, we also consider the reverse case, called "Train on Real, Test on Synthetic" (TRTS). Steps 1, 2 and 3 are identical, and steps 4 and 5 are interchanged as:

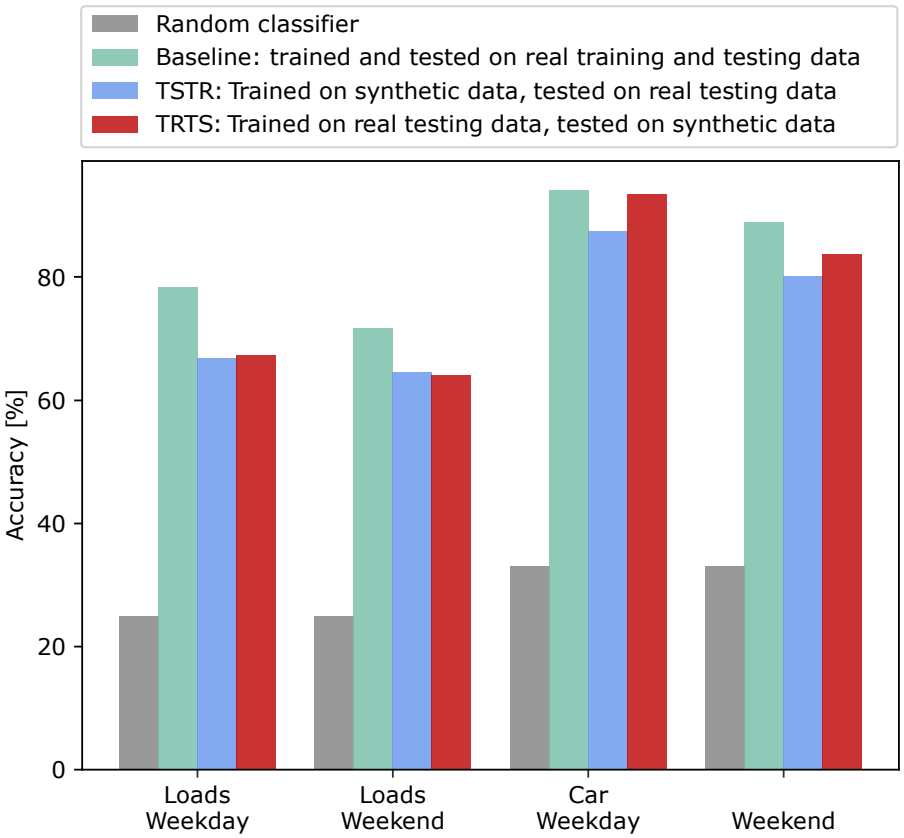


Figure 3.10: 'Train on Synthetic, Test on Real' and 'Train on Real, Test on Synthetic' accuracy scores using the trained GANs relative to random and baseline classifiers. Average accuracy over 10 repetitions.

1. Train the classifier using the held-out testing data.

2. Test the classifier model using the synthetic data.

The performance of the classifiers in the TSTR and TRTS experiments presented in Figure 3.10 shows that the synthetic data generated by the trained GANs is useful for subsequent applications.

3.3.2.9 Scaling factor transition characterisation

The unit-less normalised profiles generated by the trained GAN networks must then be scaled by a scaling factor consistent with a given home to produce profiles in energy units.

Transition matrices are used to model the probability of transitioning from one scaling factor f_t to the next one f_{t+1} in subsequent days. Using these matrices allows

the data generator to scale subsequent days of data consistently, with variability around self-correlation that matches that of real-life observed patterns for each data type and weekday type (weekday or weekend day). The space of possible scaling factors is discretised into m intervals. The probability of transitioning from discrete factor intervals i and j is then:

$$p_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,k}} \tag{3.45}$$

Where $n_{i,j}$ is the number of times a transition between intervals i and j was recorded in subsequent days of data available.

As the probability of scaling factors is not evenly distributed between the minimum and maximum factors, a non-uniform discretisation approach is adopted, based on percentile intervals, with finer data intervals for more common, lower scaling factors, and wider intervals for less common ones. This ensures that we retain granularity and information for more common lower factors. Furthermore, the 2D piecewise linear interpolation is used to fill in gaps in probability intervals, while ensuring that the sum of probabilities for the next day always equals one.

Matrices of scaling factors transition probabilities $p_f(f_{t+1} \mid f_t, \kappa_t, \kappa_{t+1})$ are illustrated in Figure 3.11.

3.3.2.10 Behaviour cluster transition characterisation

In the case of behaviour-dependent data (household loads, EV patterns), the probabilities $p_c(\kappa_{t+1} \mid \kappa_t, \nu_t, \nu_{t+1})$ of transitioning from one behaviour cluster to another in subsequent days are similarly characterised for each day type transition (for ν_t weekday or weekend day), so that profiles can be generated using the adequately trained GAN networks. Variations in generated behaviour thus match real-life patterns for each new day.

3.3.3 Data generation

From the data processing described in Section 3.3.2, the following input data for HEDGE is obtained:

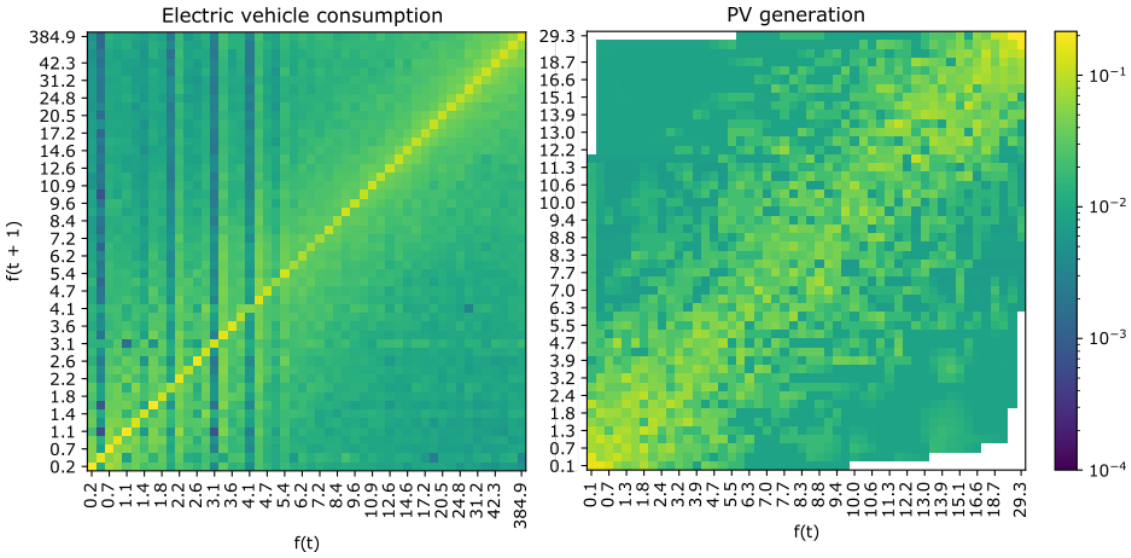


Figure 3.11: Transition probability matrices between profile scaling factors in subsequent days.

(a) Behaviour cluster transition matrices P_κ

(b) Normalised profiles generator (per data type, day type and behaviour cluster)

(c) Scaling factors transition matrices P_f

Behaviour clusters (e.g., based on Figure 3.6, to which cluster is the home closest on the day preceding the start of data generation?) and scaling factors (i.e., what is the total energy used in the day preceding the first day of data generated?) are first initialised for each home. These do not have to be real-time detailed data, but rather aim to give an indication of the type of home considered. They are automatically selected in HEDGE to match their distribution in the original dataset if not specified by the user. Then, a Markov chain mechanism uses these to generate profiles for successive days, consistent across both scaling factors and behaviour clusters. The probabilistic Markov chain transition rules are:

1. For behaviour-dependent data types, select behaviour cluster κ_t based on the behaviour cluster transition matrix $P_\kappa(\kappa_{t+1} \mid \kappa_t, \nu_t, \nu_{t+1})$, to select the appropriate GAN profile generator.

2. Generate a population of normalised profiles using pre-trained GAN weights, and randomly select one.

3. Scale the profile using a scaling factor according to the probabilities in the scaling factors transition matrices, from discrete distribution $p_f(f_{t+1} \mid f_t, \kappa_t, \kappa_{t+1})$

3.3.4 Energy user privacy preservation

The method proposed mitigates privacy concerns for experimentation with realistic residential energy data. During the data pre-processing and neural network training phase (Section 2), only anonymised disaggregated data is used from established datasets. In the data generation phase (Section 3), only pre-computed statistics and weights derived from these anonymised datasets are required. Moreover, the generated data does not pertain to any real energy user. Rather, the generated data is synthetic but realistic data that can be used for experimentation.

3.4 Other data sources and parameters

. Other data sources and parameters used by the environment are listed below. The primary contribution of this thesis is however not in the numerical results that may be input data-dependent, but rather in the coordination methodologies. Different case studies could be run with different input data and methodological insights would hold.

- The social cost of carbon is set at 70 £/tCO_2, consistent with the UK 2030 target [259].

- Weather data is taken from the instantaneous 2-dimensional hourly data collection in Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2) from NASA [260], at latitude 51.752022, longitude -1.257677, corresponding to Oxford, UK.

- The electricity real-time prices are historical data from the Agile tariff by Octopus Energy, obtained via their API [261], for Southern England. The year

2021 was selected in Chapters 4 and 5 to avoid the atypical price fluctuations observed in the subsequent year, 2022, as visually demonstrated in Figure 3.12. These fluctuations, characterised by extreme price hikes, may not provide representative patterns for demand-side response price signals in the long term. Moreover, these price dynamics in 2022 exhibited a departure from typical intra-day variability, primarily attributable to values consistently reaching the price cap. Experiments with years 2020 and 2021 were conducted in Chapter 6 to validate results.

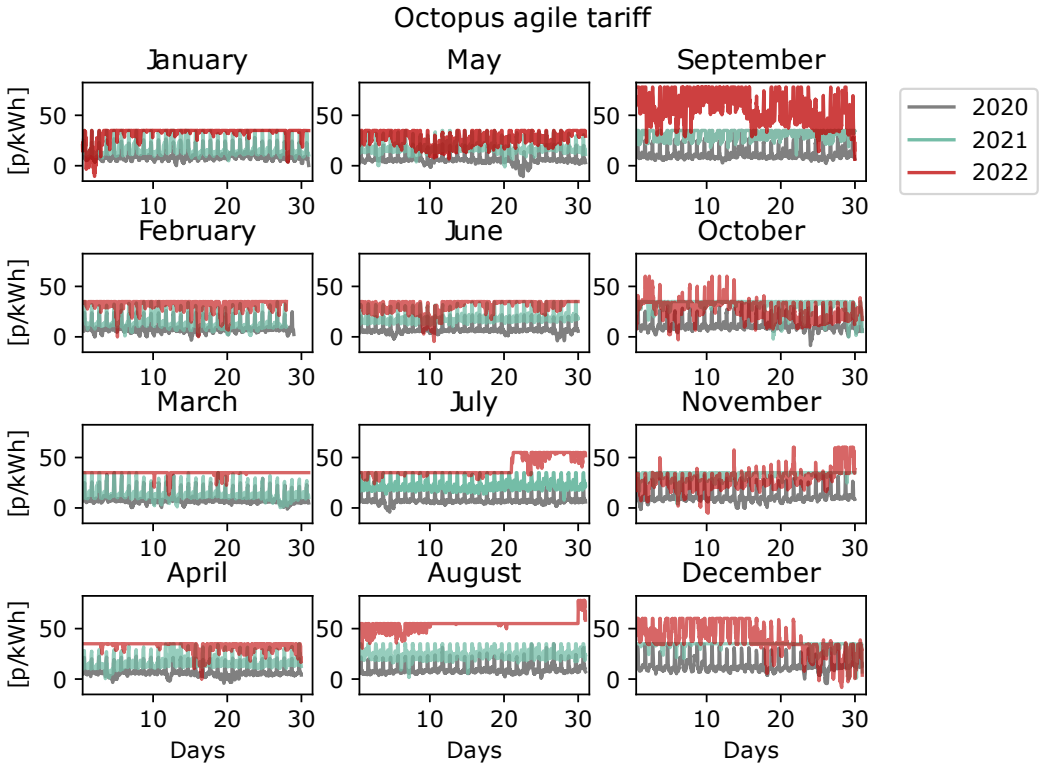


Figure 3.12: Octopus Energy Agile tariff in 2020, 2021 and 2022.

- The low solar heat gains in January are neglected [262].
- EV specifications are those of the Nissan Leaf, one of the more affordable EVs that is commonly sold in the UK and, as such, that is likely to be representative of EVs that will be widely adopted when ICE cars are phased out [263]

- The grid carbon intensity was obtained from the "Carbon Intensity API", a carbon intensity forecast tool developed by the National Grid ESO [264]

- It is assumed that 10% of loads can be deferred by up to five hours. This share is conservative, given the range of estimates in the literature: empirical work in [60] indicates 15% of loads could be flexible, a third of household loads is estimated to be wet and cold loads (assumed to be partially flexible) in [21]. The Committee on Climate Change estimates that as much as 53% of household demand could be flexible in the future [13].

- The low-voltage network model considered in this thesis is the IEEE European Test Feeder, which contains 55 homes [265].

Other model parameters are tabulated in Appendix C.

3.5 Concluding remarks

In order to answer the main research question of this thesis, this chapter starts by answering the first sub-question: *Can the efficacy of algorithms that coordinate residential energy flexibility be assessed?*.

An open-source testing and benchmarking environment for MARL algorithms was developed, applied to the problem of residential DER coordination, along with a home energy data generator (HEDGE) to generate realistic and consistent input data that MARL algorithms can use for training and evaluation. Chapter 4 now uses this environment to answer the next research sub-question of this thesis.

Criterion addressed	**Verification**
1. MARL-based implicit coordination testing environment: a testing framework specific to residential DSR coordination algorithms must be developed. It must model the local-level control of existing DERs (intermittently available EVs, space heating, household holds, PV) in response to existing energy tariffs in the form of one-way communication signals to the users. The experimenting and testing environment must include home and network modelling as well as the generation of sufficient realistic input data for robust training and testing.	✓

Table 3.3: Current assessment of algorithm success criteria set in Section 1.3

4

Optimisation-informed independent Q-learning

Contents

4.1 Introduction

The potential of MARL-based implicit cooperation strategies for residential DER coordination was identified in Chapter 2. Using the testing environment developed in Chapter 3, this chapter now investigates whether independent learners can successfully coordinate residential energy flexibility without sharing private data.

A new class of MARL-based implicit cooperation strategies for residential DSR is proposed to best use the flexibility offered by increasingly accessible assets such as PV panels, EV batteries, smart heating and flexible loads.

In a simulated rehearsal phase, agents learn RL policies using a data-based, model-free statistical approach by exploring a shared environment and interacting with decentralised partially observable Markov decision processes (Dec-POMDPs), either through random exploration or learning from convex optimisation results. They aim to learn to cooperate to reach system-wide benefits by assessing the global impact of their individual actions, searching for trade-offs between local, grid and social objectives. The pre-learned policies are then used to *implicitly* cooperate by making decisions under uncertainty given limited local information only. No further communication and only minimal, constant local computations are required to implement the pre-learned policies. As the computation burden is handled prior to implementation, this mitigates the computational challenges of large-scale real-time control.

The experiments proposed offer an understanding of the coordination mechanisms required to transition from a simple single-agent reinforcement learning methodology, Q-learning, to multi-agent coordination. Whereas standard independent learners fail to achieve coordination under partial observability in a stochastic environment, the primary novel contribution of this chapter is the introduction of *optimisation-informed* independent learning, which overcomes this challenge. Moreover, the use of marginal rewards further enhances learnability[1], so that agents can assess their individual contribution to global rewards. These contributions address the coordination performance scalability issue for agents with partial observability in a stochastic environment seeking to maximise rewards that also depend on other concurrently learning agents. However, while the fixed-size Q-tables avoid the curse of dimensionality, as the state and action space sizes remain constant with the number of agents, the optimisations and marginal reward computations represent a computational burden for training at scale. This issue will be tackled in Chapter 5.

[1] "the sensitivity of an agent's utility to its own actions as opposed to actions of others, which is often low in fully cooperative Markov games" [206]

Chapters 4 and 5 first consider the case without network management. Voltage costs are not included in the rewards and objective function (see Section 3.1.4). However, PandaPower simulations can assess the impacts of the decisions made on the network without optimising network management.

4.2 Methodology

In the MARL approach, independent prosumers learn to make individual decisions, aiming to together maximise the statistical expectation of the objective function in Section 3.1.2, as laid out in Section 3.2.

The Q-learning methodology used in this chapter is first introduced in Section 4.2.1. Then, variations on the learning method are proposed in Section 4.2.2, with different experience sources, multi-agent structures and reward definitions.

In this section, we use the notation Q for the Q-tables to reduce the notation, but note that each agent i learns its own Q_i estimate.

4.2.1 Q-Learning

This chapter first experiments with the Q-learning algorithm, a model-free, off-policy[2] RL methodology. Its simplicity and proof of convergence make it suited to developing novel learning methodologies in newly defined environments [9]. Moreover, previous work has shown that IQL can achieve reasonable performance in different multi-agent tasks with 2 to 10 agents [223].

The states and actions are discretised into intervals. State-actions values $Q(s, a)$ then represent the expected value of all future rewards $r_t \; \forall \; t \in \mathcal{T}$ when taking action a in state s according to policy π:

$$Q(s, a) \triangleq E^{\pi}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} ... | s_t = s, a_t = a] \tag{4.1}$$

where γ is the discount factor setting the relative importance of future rewards. Estimates are refined incrementally as

$$\hat{Q}(s, a) \leftarrow \hat{Q}(s, a) + \alpha\delta \tag{4.2}$$

[2]Refer back to Section 2.2.2 for definitions of these terms.

where δ is the temporal-difference (TD) error,

$$\delta = \left(r_t + \gamma \hat{V}(s^{\text{next}}) - \hat{Q}(s, a)\right) \tag{4.3}$$

$\hat{V}$ is the state-value function estimate,

$$\hat{V}(s) = \max_{a^* \in \mathcal{A}(s)} \hat{Q}(s, a^*) \tag{4.4}$$

and α is the learning rate.

Agents follow an ϵ-greedy policy [185] to balance exploration of different state-action pairs and knowledge exploitation. The greedy action with the highest estimated rewards is selected with probability $1 - \epsilon$ and random actions otherwise.

$$a^* = \begin{cases} \arg\max_{a^* \in \mathcal{A}} \hat{Q}(s, a^*) & \text{if } x \sim U(0,1) > \epsilon \\ a \sim p(a) = \frac{1}{|\mathcal{A}|} \ \forall \ a \ \in \mathcal{A} & \text{otherwise} \end{cases} \tag{4.5}$$

Henceforth, the estimates $\hat{Q}$ and $\hat{V}$ are referred to as Q and V to reduce the amount of notation.

4.2.2 Variations of the learning method

Different experience sources and reward definitions are proposed within the MARL approach. The performance of these combinations of algorithmic possibilities will be assessed in Section 4.3 to inform effective model design.

4.2.2.1 Experience sources

In data-driven strategies, the source of collected experience is crucial as it directly determines learning. We propose two such sources of collecting experience:

- **Environment exploration**. Traditionally, agents collect experience by interacting with an environment [185].
- **Optimisations**. A novel approach collects experience from optimisations. Learning from entities with more knowledge or using knowledge more effectively than randomly exploring agents has previously been proposed, as

with agents "mimicking" humans playing video games [266]. Similarly, agents learn from convex "omniscient" optimisations on historical data with perfect knowledge of current and future variables. This experience is then used under partial observability and control for stable coordination between prosumers at scale. Note in this case that, although the MARL learning and implementation are model-free, a system model is used to run the convex optimisation and produce experience from which to learn. A standard convex optimiser uses the same data that would be used to populate the environment explorations but solves over the whole day-horizon with perfect knowledge of all variables using the problem description in Section 3.1. Then, at each time step, the system variables are translated into equivalent RL $\{s_t, a_t, r_t, s_{t+1}\}$ tuples for each agent[3], which are used to update the policies in the same way as for standard Q-learning.

4.2.2.2 Reward definitions

The reward definition is central to learning as its maximisation forms the basis for incrementally altering the policy [185].

This chapter defines the global reward as the negative sum of grid, distribution and battery costs, defined in Section 3.1.2.

$$r^t = -(c_{\mathrm{g}}^t + c_{\mathrm{d}}^t + c_{\mathrm{b}}^t) \tag{4.6}$$

Four variations of the Q-table update rule are proposed for each experience step tuple collected $(s_i^t, a_i^t, r^t, s_i^{t+1})$, building on the main update rule defined as:

$$Q(s_i^t, a_i^t) \leftarrow Q(s_i^t, a_i^t) + \alpha\delta \tag{4.7}$$

- **Total reward**. The instantaneous total system reward $r^t = \hat{F}_t$ is used to update the Q-table Q^0.

$$\delta = r^t + \gamma V^0(s_i^{t+1}) - Q^0(s_i^t, a_i^t) \tag{4.8}$$

[3]At time steps with no flexibility due to the environment and user constraints, a random action was selected to avoid bias

- **Marginal reward**. The difference in total instant rewards r^t between that if agent i selects the greedy action and that if it selects the default action is used to update Q^{diff}. The default action a_{default} corresponds to $a = 1$, where no flexibility is used. The default reward $r^t_{a_i=a_{\text{default}}}$, where all agents perform their greedy action apart from agent i, which performs the default action, is obtained by an additional simulation[4]

$$\delta = \left(r^t - r^t_{a_i=a_{\text{default}}}\right) + \gamma V^{\text{diff}}(s_i^{t+1}) - Q^{\text{diff}}(s_i^t, a_i^t) \tag{4.9}$$

- **Advantage reward**. The difference between Q^0 values when i performs the greedy and the default action is used. This corresponds to the estimated increase in rewards not just instantaneously but over all future states. No additional simulations are required as the Q-table values are refined over the normal course of explorations.

$$\delta = \left(Q^0(s_i^t, a_i^t) - Q^0(s_i^t, a_{a_i=a_{\text{default}}})\right) - Q^{\text{A}}(s_i^t, a_i^t) \tag{4.10}$$

4.2.3 Parameter tuning

Learning hyper-parameters were tuned using an iterative approach.

Although a highly time-consuming step in the methodological development process, this process was crucial to ensure that the assessment of a given MARL methodology was meaningful. Results can thus be presented with values that optimise the performance of the MARL algorithm in the problem at play, as opposed

[4]Note that, contrary to the Shapley value, the aim here is not to guarantee that the sum of individual marginal rewards is exactly equal to the value created by the grand coalition relative to the baseline (referred to as the *efficiency* axiom of a payoff allocation) [76]. As stated in the scope in Section 1.4, the fair allocation of the coalition's value amongst its members is beyond the scope of this thesis, which focuses on the operational problem. Here, the aim is merely to compute useful reward signals to learn policies that increase the likelihood of actions that are beneficial for the system. In value-based reinforcement learning, the optimal policy is found indirectly by finding the optimal value function (see Section 2.2.2.5). As stated in [194]: "Value functions define a partial ordering over policies. A policy π is defined to be better than or equal to a policy π' if its expected return is greater than or equal to that of π' for all states. In other words, $\pi \geq \pi'$ if and only if [the value function] $v_\pi(s) \geq v_{\pi'}(s)$ for all $s \in \mathcal{S}$". Therefore, here it only matters that the rewards result in an ordering of action value functions consistent with the expected long-term rewards yielded by the actions for the system. So long as this ordering is maintained, there is no need for guarantees that the sum of the marginal rewards over for all agents is exactly equal to the total value of the coalition.

to results contingent on an arbitrary hyper-parameter selection. The tuning process, therefore, required careful experimentation, analysis, and adjustment.

The following steps were followed to select optimal hyper-parameter values:

1. Defining the performance metric: here, hyper-parameters are sought that maximise savings in terms of the global rewards (as defined in Section 4.2.2.2) relative to the baseline. Specifically, the aim is to maximise the *best* savings obtained amongst the possible learning methodologies defined in the previous sub-sections.

2. Choosing hyper-parameters to tune: hyper-parameters are identified that have a significant impact on the RL algorithm's performance.

3. Defining the initial set of hyper-parameter values and the search space: values are initialised, and a search space is specified for each hyper-parameter, which includes the range or set of values that will be explored during the tuning process. Meaningful and realistic search spaces are based on prior knowledge.

4. Iterative search: for each parameter in the list:

   (a) Evaluate the savings for a range of values over multiple learning trajectories to mitigate the impact of random numbers and input data variation. Identify the best hyper-parameter value.

   (b) If this value differs from the default, iterate and refine: update the default hyper-parameter value, identify promising hyperparameter configurations and gain insights into their effects. Update the search space for subsequent iterations, focusing on the regions that have shown better performance. Go back to (4) to loop through all hyper-parameters again.

   (c) If this is the same default value, continue looping through the list of hyper-parameters.

4.3 Results

The performance of the residential flexibility coordination strategies presented in Section 4.2 are compared to baseline and upper bound scenarios for increasing numbers of homes. We use the acronyms in Table 4.1 to refer to the possible methodological combinations of experience source and reward definitions.

	Environment explorations	Optimisation-informed
Total rewards	TE	TO
Marginal rewards	ME	MO
Advantage rewards	AE	AO

Table 4.1: Acronyms used to refer to independent Q-learning methodological variations

Results show that only the algorithms learning from optimisations maintain stable coordination performance at scale, while the performance of standard independent learning algorithms (TE) drops in this context of stochasticity and partial observation. The optimisation-based algorithm which uses marginal rewards (MO) performs best. The following subsections further elaborate on these results.

Section 4.3.1 first presents the experimental set-up, and Section 4.3.2 the hyper-parameter selection. Section 4.3.3 then analyses the coordination performance of independent learners with environment exploration and with optimisation-informed learning as the number of homes increases. The implementation of policies at the home level is illustrated in Section 4.3.4. Subsequently, the reliability of the performance of algorithms is analysed using a multi-criteria analysis in Section 4.3.5. Finally, while the methodologies developed in this chapter achieve coordination scalability, remaining limitations in terms of computational scalability are identified in Section 4.3.6.

4.3.1 Set-up

Learning experiments are performed over 20 epochs consisting of an exploration, an update and an evaluation phase. First, the environment is explored over two training episodes of duration $|\mathcal{T}| = 24$ hours. Exploration data is generated by HEDGE, and

optimisations are performed on this input data in the case of optimisation-informed learning. Learning in batches of multiple episodes helps stabilise learning in the stochastic environment. Then, Q-tables are updated based on the rules presented in Section 4.2.2. Finally, new testing data is generated (without optimisations), and an evaluation is performed using the trained deterministic greedy policy. Ten repetitions are performed to assess the learning over different trajectories, thereby mitigating any potential bias introduced by the initial seed values. This number of repetitions across different seeds is recommended in [226] for the reliable standardised performance evaluation protocol for cooperative MARL, and frequently selected in the literature [157, 215, 224, 256, 267, 268]. Figure 4.3 shows that the spread of average rewards over this number of repetitions is is adequately small, such that meaningful conclusions about the relative performance of the different methodologies to be drawn, beyond the effect of randomness of individual learning trajectories.

Case study results using different experience sources, reward definitions and MARL structures are presented in Figure 4.4. Positive values denote savings relative to a baseline scenario where all agents are passive, i.e. not using their flexibility with EVs charged immediately and no flexible loads delayed. In the first epoch, rewards are below the baseline. As agents collect experience and update their policies at each epoch, improved policies are learned, some of which can outperform the baseline. An upper bound is provided by results from "omniscient and time-travelling" convex optimisations, which are, however, not achievable in practice for three main reasons. Firstly, they use perfect knowledge of all the environment variables in the present and future, despite uncertainty in renewable generation, mix of the grid, and customer behaviour. Optimisation with inaccurate data would lead to suboptimal results. Secondly, prosumers may not be willing to yield their data and direct control to an external entity. Finally, central optimisations become computationally expensive for real-time control of large numbers of prosumers.

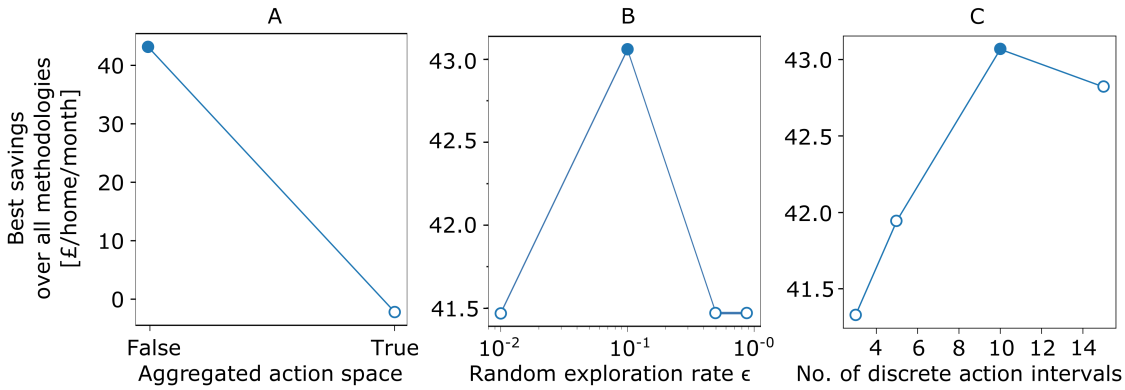


Figure 4.1: Parameter tuning for independent Q-learning. Savings obtained per home and month, only varying one hyper-parameter at at a time. The highest savings are indicated by the full dot.

4.3.2 Parameter tuning

All sensitivity analyses from the parameter tuning procedure presented in Section 4.2.3 are illustrated in Figures 4.1 and 4.2. Two principal insights drawn from these findings are:

1. A disaggregated action space provides superior performance, with three action variables for household consumption, heating and EV charging $a^t_{\text{cons},i}$, $a^t_{\text{heat},i}$ and $a^t_{\text{bat},i}$ (Figure 4.1-A).

2. Q-learning performs best with a small state space (one observation) (Figure 4.2-C). The grid price C^t_{g}, is the most useful observation, i.e. the sum of the grid electricity price and the product of the carbon intensity of the generation mix at time t and the social cost of carbon. This state space is therefore adopted in this chapter.

The key parameters selected as a result of the hyper-parameter tuning are tabulated in Table 4.2.

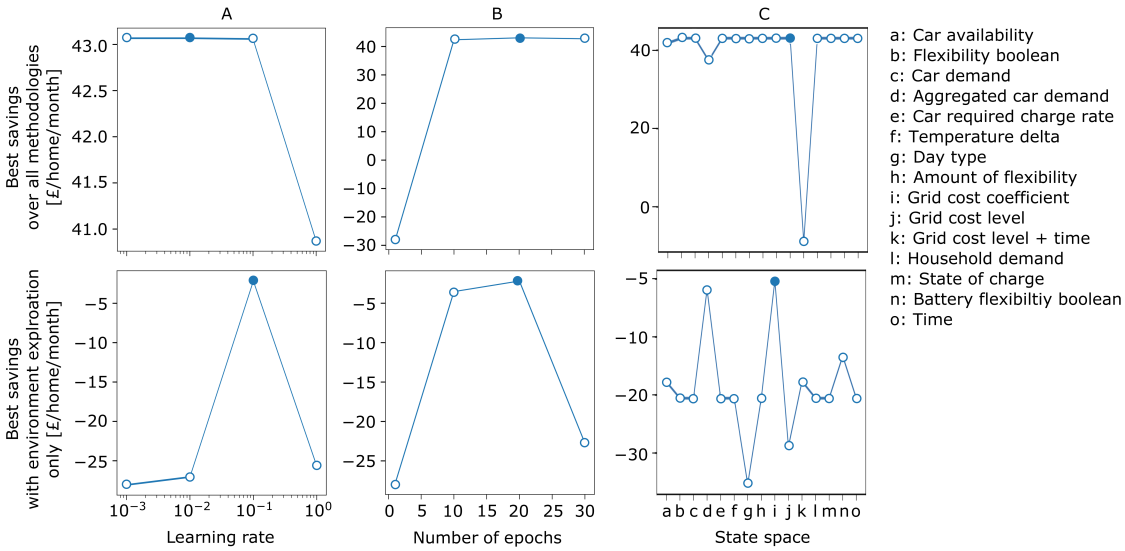


Figure 4.2: Parameter tuning for independent Q-learning. Savings obtained per home and month, only varying one hyper-parameter at at a time. The highest savings are indicated by the full dot. If the best hyper-parameter value cannot be determined conclusively by examining the best methodology's savings only, the hyper-parameter is selected amongst the set of values maximising the best method based on the environment exploration-informed learning savings.

Description	**Value**
Depreciation rate	$\gamma = 0.9$
Learning rate	$\alpha_0 = 1 \times 10^{-1}$
Exploration rate	$\epsilon = 0.1$
Number of epochs	20
Number of intervals in discretised action space	10

Table 4.2: Independent Q-learning learning parameters

Moreover, we confirm that ten repetitions are adequate to draw conclusions about the relative performance of different methodologies in Figure 4.3. The influence of initial random seeds may introduce inconsistencies in the ordering of averages when based on one or five repetitions. However, by averaging over ten repetitions, we can draw meaningful and consistent conclusions, evident in non-overlapping average reward spreads for all plotted averages over 30 repetitions. This reaffirms the appropriateness of choosing ten repetitions as a hyperparameter, aligning with recommendations in the literature (see Section 4.3.1).

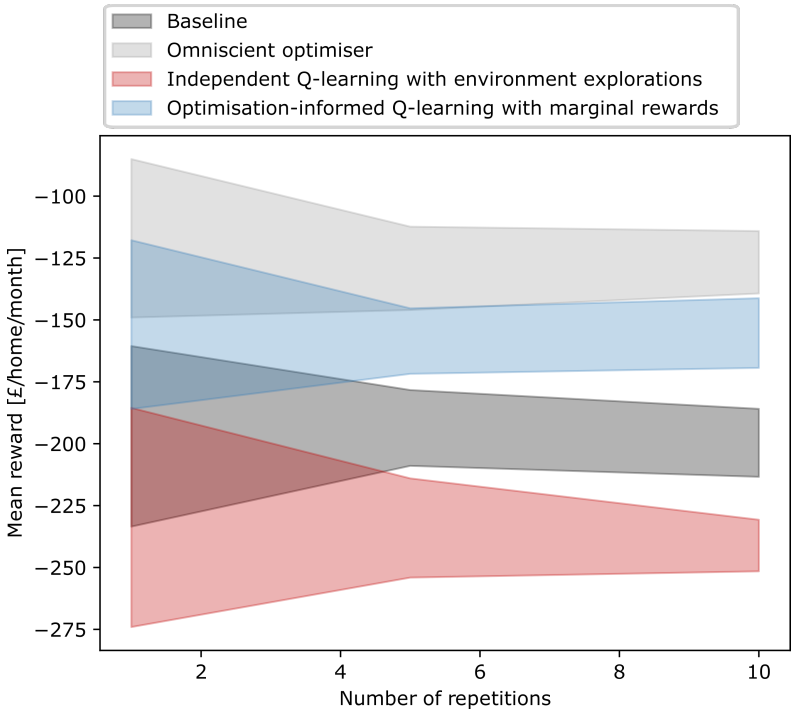


Figure 4.3: The shaded area represents the spread between the minimum and maximum average rewards of key methodologies for 30 different averages over different numbers of repetitions with five homes.

4.3.3 Environment exploration and optimisation-based learning

Environment exploration-based learning Figure 4.4 shows that environment exploration-based MARL using total rewards (full red line), the standard MARL framework, achieves savings just above the baseline for a single agent. The savings then drop as the number of cooperating agents increases, reaching a minimum of £43.47 per home and month worse than the inflexible baseline for four homes. Without a coordination mechanism, coordination challenges arise for independent learners under partial observability in a stochastic environment [206]:

1. The incompatibility of individual policy equilibriums with a global Pareto optimal,

2. The non-stationarity of the environment due to the evolving on-policy behaviour of other concurrently learning agents affects convergence, and

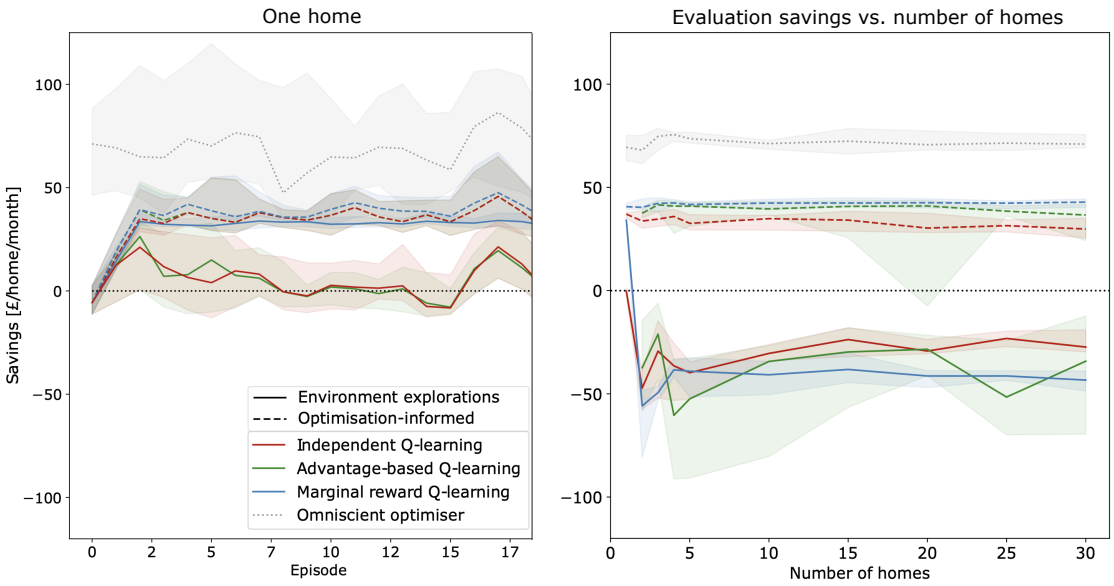


Figure 4.4: The left-hand side plot shows the five-epoch moving average of evaluation rewards relative to baseline rewards (zero value) for a single prosumer. The right-hand side plot shows the mean of the ten evaluations (after training is over) against the number of prosumers. Lines show median values and shaded areas the 25th and 75th percentiles over the ten repetitions. The performance of the standard MARL algorithm with total rewards and environment explorations (TE, full red line) drops as the number of concurrently learning agents in the stochastic environment increases; the best-performing alternative algorithm proposed (dotted blue) maintains high performance at scale.

3. The stochasticity of the environment prevents agents from discriminating between their own contribution to global rewards and noise from other exploring agents or the environment.

Using advantage rewards (AE), based on estimates of the long-term value of actions relative to that of the baseline action, does not provide an improvement on the median values and increases the variability of the results. As AE uses the total reward Q^0-table as an intermediary step, its value can only be as accurate as the Q^0 estimate. Moreover, as it performs differences between the value of the global action space relative to when each agent takes the baseline action, the exploration should cover a wide range of actions, even those seldom explored combinations of greedy and non-greedy actions.

Using marginal rewards (ME), the value of each agent's action relative to the baseline action is singled out immediately by an additional simulation and used

as a reward at each time step. This allows a single agent to achieve high levels of savings relative to the inflexible baseline, as it can accurately assess the impact of its actions on the reward, providing an improvement relative to TE. Still, the performance declines as the number of agents increases, similarly to the other environment exploration-informed learning strategies.

Optimisation-informed learning Comparing trajectories in Figure 4.4, learning from the experience obtained from an optimiser consistently improves savings relative to learning from random explorations in the experiments. Consistent with the results from Section 4.3.3, MO yield superior performance relative to advantage-based learning (AO), which itself is superior to total rewards. By observing an omniscient optimiser, which takes precise decisions thanks to its perfect knowledge of all current and future system variables, the RL algorithms are able to learn policies to be used under partial observability, aiming for actions that statistically perform well under uncertainty.

MO offers the highest savings as the additional baseline simulations are best able to isolate the contribution of individual actions from variations caused by both the environment and other agents. When increasing the number of agents, the strategy can learn from optimal, stable, consistently behaving agents. Savings of £44.88 per home per month are obtained on average for 30 agents, corresponding to an 18.6% reduction from baseline costs, or 56.4% of the maximum total savings achieved by the omniscient optimiser on average.

The strategy exhibits stable performance at scale, though converging to a different type of policy in comparison with the optimiser. The MO policy saves more by smoothing out the charging and distribution grid utilisation profiles, despite smaller savings in imports and emissions costs. In contrast, the optimiser derives a larger advantage from performing arbitrage between the grid price differentials, though with higher battery and distribution grid costs. Examples of how individual home energy management varies based on the controller are illustrated in Section 4.3.4.

4.3.4 Commented illustrative day

Here, we look in detail at the actions selected by an agent coordinated using different MARL strategies. This illustrates how example RL actions translate into local energy management system behaviour. Note however that the MARL algorithms aim to generate statistically favourable outcomes when averaged over longer durations and over a larger number of agents. As such, while the average outcomes are predictable, this individual case is not meant to be generally representative but rather simply an example day in a stochastic environment.

Figure 4.5 shows an example of an evaluation day during which the final policies learned are used deterministically on a day-long batch of data. Three different policies are compared: Baseline, optimal, and MO.

The baseline and optimal act as reference points, while the latter policy has been identified in Section 4.3.3 as the highest-performing scalable policy when the number of agents increases.

Subplot (a) shows the wholesale prices and the grid carbon intensity for the example day, as well as the resulting grid cost coefficient C_g given a social cost of carbon of 70 £/tCO_2. This coefficient informs the choice of action.

Subplot (b) shows the EV at-home availability and consumption. Enough charge must be available at the start of trips, and the car can not be charged while on a trip. Comparing the battery level profiles in this example day, in the baseline, the EV is charged as soon as it is plugged in, given battery capacity and charging rate constraints. The optimiser policy sells energy from the battery when prices increase to take advantage of the price differentials, whereas the MO policy flattens out the charging profile.

Subplot (c) shows cumulative rewards over time for each policy. An interplay is thus illustrated between the costs of battery depreciation and distribution network congestion on the one hand, and the opportunity for energy arbitrage to save on grid energy and emissions costs on the other. While the MO policy saves more by smoothing out the charging and distribution grid utilisation profiles, the optimiser derives a larger advantage from the grid price differentials in grid imports,

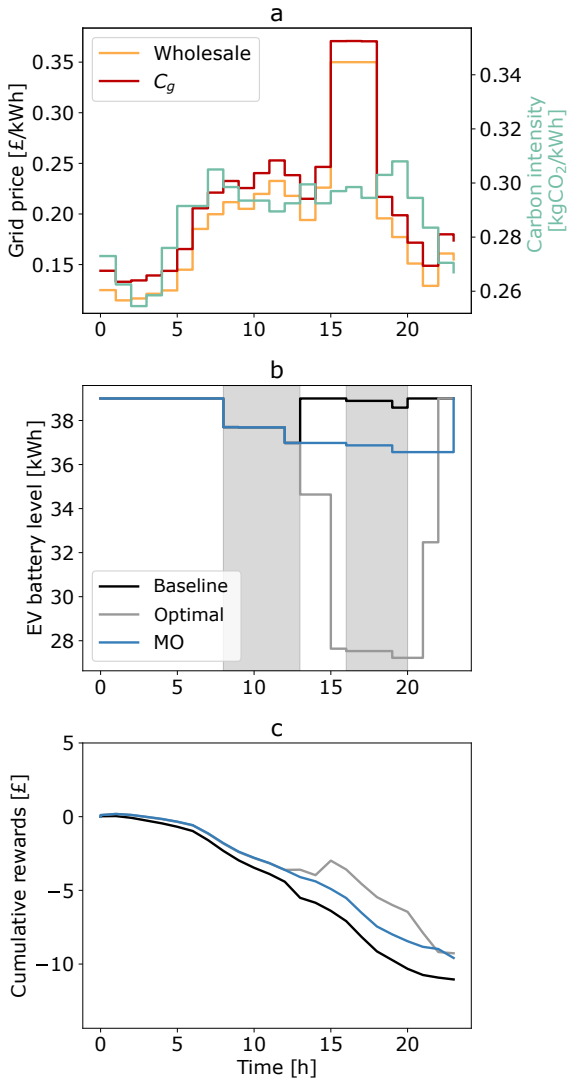


Figure 4.5: Example of local home energy system variables for the policy learned using optimisation-based learning using marginal rewards (MO), compared with the inflexible baseline and omniscient and time-travelling centralised optimiser control actions

though with higher battery and distribution grid costs. The weight applied to each of those competing objectives in the objective function will have a direct impact on the policies that are learned.

4.3.5 Reliability

The performance of RL algorithms is assessed not only in terms of average performance, but also in terms of variability and risk. This is especially critical for RL algorithms interacting with critical infrastructure such as homes and the

electricity network.

We use the metrics proposed in [269], which measure both:

- The dispersion of results: the width of a distribution, using the Inter-quartile range (IQR); and
- The risk of algorithms: to assess worst-case scenarios, the metrics measure the heaviness and extent of the lower tail of the distribution using the Conditional Value at Risk (CVaR), also known as "expected shortfall". Here, the expected loss is measured for the worst-case scenarios, over the lower quantile $\alpha = 0.05$.

Moreover, it measures these over two axes:

- Across time: smooth monotonic improvement is preferable to noisy fluctuations around a positive trend, or unpredictable swings in performance,
- Across runs: to measure sensitivity to both stochasticity from the environment and stochasticity from the training procedure.

The 5 metrics utilised are thus tabulated in Table 4.3 and listed below:

1. Dispersion across Time (DT): IQR across Time. This is used to isolate higher-frequency variability, rather than capturing longer-term trends. For detrending, differencing is used $y'_t = y_t - y_{t-1}$.
2. Short-term Risk across Time (SRT): CVaR on Differences. This measures the most extreme short-term drop over time, i.e. the CVaR to the changes in performance from one evaluation point to the next.
3. Long-term Risk across Time (LRT): CVaR on Drawdown. This measures the potential of an algorithm to lose a lot of performance relative to its peak, even if on a longer timescale. The Drawdown at time t is the drop in performance relative to the highest peak so far.

4. Dispersion across Runs (DR): the IQR across training runs at a set of evaluation points. A low-pass filtering of the training data is first performed, to filter out high-frequency variability within runs (this is instead measured using Dispersion across Time, DT).

5. Risk across Runs (RR): CVaR across Runs. This is the CVaR of the final performance of all the training runs.

	Dispersion (D)	Risk (R)
Across time (T) (within training runs)	DT: IQR within windows, after detrending	Short-term (SRT): CVaR on first-order differences Long-term (LRT): CVaR on Drawdown
Across Runs (R)	DR: IQR across training runs, after low-pass filtering	RR: CVaR across runs

Table 4.3: Reliability metrics from [269]

Figure 4.6 shows the metrics corresponding to the experiments for 30 homes. It is found that optimisation-informed learning not only increases average savings, but also reduces both variability and risk across all metrics. It even offers an improvement over a centralised convex optimiser in terms of Dispersion across Time (DT), Short-term Risk across Time (SRT) and Long-term Risk across Time (LRT).

Using marginal rewards increases the dispersion across runs (DR) and dispersion across time (DT). However, the commensurate cost savings warrant these augmentations, which remain lower than the inflexible baseline.

4.3.6 Computational scalability

Finally, we analyse the computational scalability of the proposed optimisation-informed learning approach with marginal reward computation in Figure 4.7.

As performed on an Intel(R) Core(TM) i7-9800X CPU @ 3.80GHz, training times for one learning trajectory are 1 minute 7 seconds for one agent, and 41 minutes 7 seconds for 30 agents, excluding testing. Although the policy can then be directly applied at the household level during operation with minimal computational

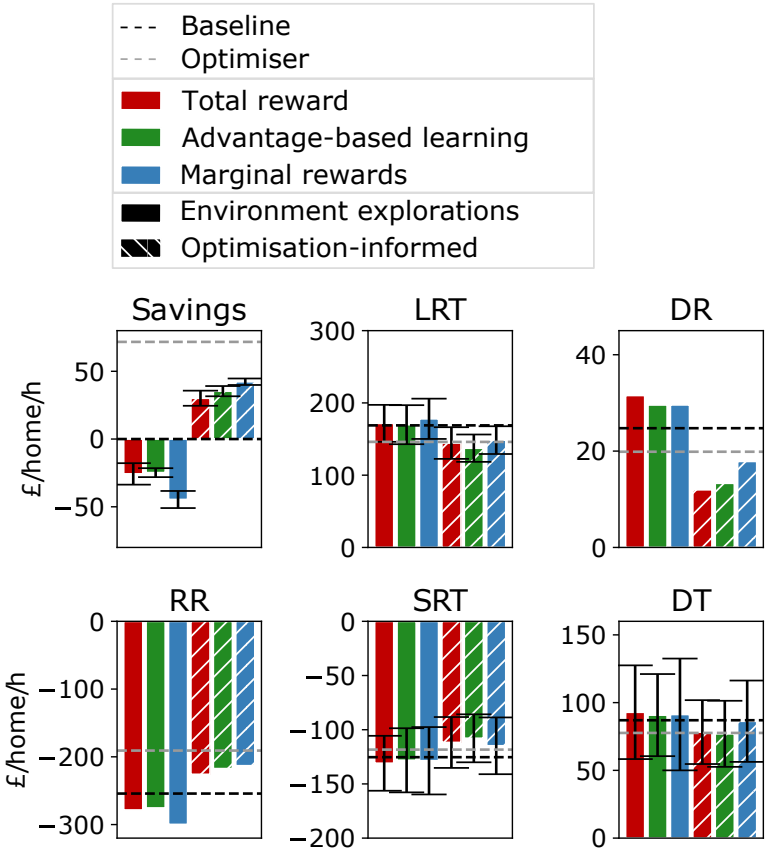


Figure 4.6: Reliability metrics: variations on independent Q-learning in experiments with 30 homes. DT: dispersion across time. DR: dispersion across runs. SRT: short-term risk across time. LRT: long-term risk across time. RR: risk across runs.

burden (see Section 3.2.4), Figure 4.7 shows the computational burden for training itself grows with second-order polynomial time $O(n^2)$ as the system size increases. This is due to the need to run optimisations on input data, and to perform additional simulations for each agent to compute marginal rewards.

4.4 Concluding remarks

As discussed in Section 2.2.3, a fundamental challenge in MARL is the trade-off between fully centralised value functions, which are impractical for more than a handful of agents, and the more straightforward independent learning of individual action-value functions. However, while IQL can sometimes perform well in multi-agent task [223], an ongoing issue with the latter approach has been that of

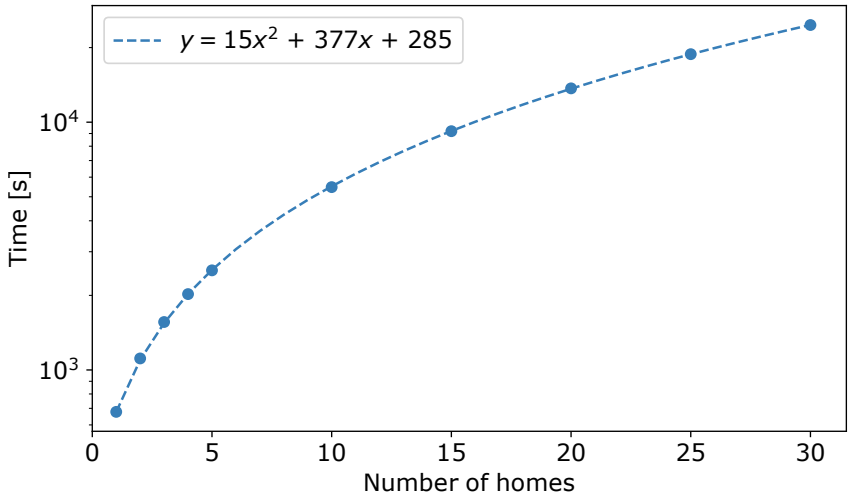


Figure 4.7: Training time required for independent Q-learning with optimisations and marginal rewards as the number of homes in the system increases (10 repetitions over the whole experiments). The line is fitted through data points using SciPy's optimize.curve_fit non-linear least squares function [270]). Plotted in logarithmic scale.

convergence at scale, as agents do not have explicit representations of interactions between agents, and each agent's learning is confounded by the learning and exploration of others [271]. As shown in Figure 4.4, the Pareto selection, non-stationarity and stochasticity issues presented in Section 2.2.3.2 have prevented environment exploration-based learners from achieving successful MARL cooperation at scale for agents under partial observability in a stochastic environment.

Thus, the case study presented in this chapter shows that the coordination performance of agents in the standard IQL framework plummets for increasing numbers of agents. Therefore, an additional coordination mechanism is required for the coordination of agents under partial observability in the stochastic environment used in this thesis.

The novel combination of marginal rewards and optimisation-informed learning proposed in this chapter offers significant improvements on these scalability and convergence issues. On the one hand, marginal rewards help improves learnability as agents isolate their marginal contribution to total rewards. Assessing the impact of individual actions on global rewards accurately is key to the effective coordination of a large number of prosumers. On the other, learning from the

results of convex optimisations allows agents to learn successful policy equilibriums from omniscient, stable, and consistent solutions. Allowing agents to learn from omniscient, stable, and consistent optimisation solutions can successfully act as an equilibrium-selection mechanism.

Overall, the new class of optimisation-based learning performs significantly better across different numbers of homes, with higher savings and lower inter-quartile range than environment-based learning at scale. Moreover, comparing the MARL policies with the actions performed by an optimiser has shown that the problem of finding a global minimum given global information is available is different from finding a robust policy given only partial information is available, for which different methodologies are required. Finally, this method tackles acceptability issues, with no interference in personal comfort or communication of personal data. Only one-way communication is required, and there is minimal computational burden during implementation. Coordination scalability is achieved, with more than half of the optimal performance maintained even as the number of homes increases. This meets the second success criterion set in Chapter 1 and tabulated in Table 4.4.

Although computational time for pre-learning is not strictly a limiting factor as it is performed off-line ahead of implementation, this superior performance requires computational resources to run optimisations on historical data, and to perform baseline simulations to compute marginal rewards. Therefore, Chapter 5 uses the insights gained in this chapter and develops new methodologies to improve the computational scalability of the coordination further.

Criterion addressed	**Verification**
2. Coordination scalability without sharing private data: Implement cooperative multi-agent coordination algorithms in a decentralised way without sharing private data. Limit the control of appliances to the local level, with no communication of personal data, thermal discomfort, or hindrance/delay of activities, without relying on accurate individual forecasts or real-time central computations. Despite this information and control gap, cooperative multi-agent coordination algorithms implemented in a decentralised way without sharing private data should yield value relative to an uncoordinated system. This coordination performance should be maintained as the system size increases, measured in savings in energy, network and carbon costs obtained per home and month.	56.4% of optimal savings achieved for 30 homes ✓

Table 4.4: Current assessment of algorithm success criteria set in Section 1.3

5

Deep multi-agent reinforcement learning with factored critic for scalable coordination

Contents

5.1 Introduction

As part of the main research question set out in this thesis, this chapter aims to answer the sub-question: *Can coordination of residential energy without sharing private data be achieved in a computationally scalable manner?*

The optimisation-informed IQL algorithms developed in Chapter 4 achieved high coordination performance at scale for agents under partial observability in a stochastic environment, with no sharing of personal data required during execution. However, the computational scalability of the training phase remains a challenge, as it imposes an ever greater training computational burden as the size of the system increases.

This chapter, therefore, develops an alternative coordination approach and uses a deep MARL algorithm that can fully exploit the benefits of the *centralised training with decentralised execution* (CTDE) [208] paradigm, such that optimisations would no longer be required. As stated in Chapter 2, deep RL can improve efficiency and performance in problems with high dimensional, arbitrarily large and possibly continuous state spaces. In CTDE, due to partial observability or communication constraints, each agent must learn a *decentralised* policy conditioned only on local observations. However, the training itself can be *centralised* in a simulated environment, with access to additional information about the environment (e.g. global state) and other agents. Three common classes of cooperative MARL approaches used to solve a Dec-POMDP are summarised in Section 2.2.3.

Particularly, the FACMAC algorithm [222] is identified as having the potential to achieve scalable DER coordination, as it combines the advantages of both the *centralised multi-agent policy gradient* and the *value function factorisation* frameworks. FACMAC is a deep multi-agent actor-critic method that learns a single *centralised but factored* critic[1] to rehearse coordination ahead of execution. The critic factors the centralised action-value function Q_{tot} as a non-linear monotonic function of individual action-value functions Q_i.

Whilst this could address the pitfalls of both the poor coordination performance of independent learners and the intractability at scale of optimisations or centralised critic estimation, this potential has thus far not been investigated for the coordination of residential DER coordination. Here, the neural network structure and state and action space had to be optimised for the specific DER coordination problem,

[1]For the remainder of this chapter, "FACMAC" and "centralised but factored critic approach" will be used interchangeably.

including using convolutional neural networks to set up control actions for the whole day ahead. Furthermore, a supervised loss approach is adapted and extended from a single-agent RL to a MARL framework, to combine computationally "expensive" training data obtained from system optimisations results (see Chapter 4) and "cheap" environment exploration to guide the learning of multi-agent cooperation.

Results show that coordination is achieved at scale, with minimal information and communication infrastructure requirements, no interference with daily activities, and privacy protection. Significant savings are obtained for energy users, the distribution network and greenhouse gas emissions. Moreover, training times are 34 times shorter than for optimisation-informed IQL for 30 homes.

This chapter uses the environment in Chapter 3 without network management, as in Chapter 4.

5.2 Methodology

5.2.1 MARL set-up

Let us consider a fully cooperative multi-agent task, in which a team of agents interacts with the same environment to achieve some common goal, which can be modelled as a Dec-POMDP [272]. The Dec-POMDP consists of a tuple $G = \langle \mathcal{N}, S, A, P, r, O, \gamma \rangle$. Here $\mathcal{N} \equiv \{1, ..., n\}$ denotes the finite set of agents, and $s \in \mathcal{S}$ describes the true state of the environment. At each time step, each agent $i \in \mathcal{N}$ selects a continuous action $a_i \in \mathcal{A}$, forming a joint action $\boldsymbol{a} \in \boldsymbol{A} \equiv A^n$. The environment then produces a transition to the next state s' according to the state transition function $P(s'|s, a) : S \times A \times S \to [0, 1]$ and a team reward $r(s, a)$. Due to the partial observability, each agent $i \in \mathcal{N}$ draws an individual partial observation $o_i \in \mathcal{O}$ from the observation kernel $O(s, i)$. Each agent learns a policy $\pi_i(\tau_i)$, conditioned only on its local action-observation history $\tau_i \in T \equiv (\mathcal{O} \times A)^*$, which may be stochastic or deterministic. The joint policy π induces a joint action-value function: $Q^\pi(s_t, a_t) = \mathbb{E}_{s_{t+1:\infty}, a_{t+1:\infty}}[G_t|s_t, a_t]$, where $G_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i}$ is the discounted return, with $\gamma \in [0, 1)$ a discount factor.

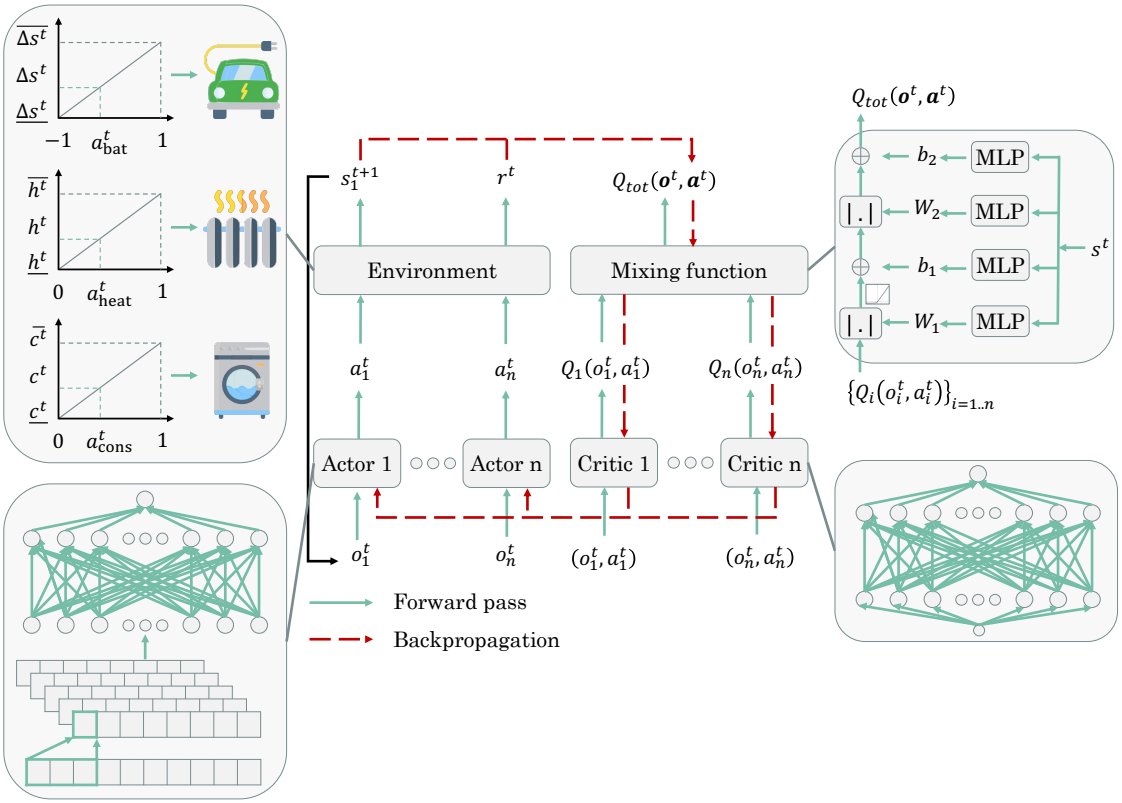


Figure 5.1: FACMAC architecture. For each agent i, one actor network selects an individual action a_i^t based on local observation o_i^t. This action is input into the local energy environment, where RL actions are translated into environment variables for household flexible loads, electric heating and EV battery charging and discharging. Based on the environment step, the next state s_i^{t+1} and reward r^t are obtained. Each agent i also estimates a critic network that estimates the individual action-value function Q_i, which is then combined into the centralised action-value function Q_{tot} via a non-linear monotonic mixing function approximated by a mixing network with non-negative weights. The actor network includes a convolutional layer followed by two hidden layers. The critic network is a linear network with one hidden layer. Only the feed-forward flow through the actor network from observation to action is needed during implementation.

5.2.2 Centralised but factored critic

Like multi-agent deep deterministic gradients (MADDPG) [215], a popular multi-agent actor-critic method, FACMAC uses deep deterministic policy gradients to learn policies. Additionally, it learns a centralised but factored critic, which combines per-agent utilities into the centralised action-value function via a non-linear monotonic function, as in QMIX [217], to improve the scalability of the centralised critic estimation. Moreover, FACMAC uses a new centralised policy gradient estimator

that optimises over the entire joint action space to allow for more coordinated policy changes and fully reap the benefits of a centralised critic.

Figure 5.1 schematically illustrates the modified factored multi-agent centralised policy gradients (FACMAC) [222] structure. For each agent i, one actor network selects an individual action a_i^t based on local observation o_i^t. Here, the actor network includes a convolutional layer followed by two hidden layers. Actions a_i^t are input into the local energy environment, where RL actions are translated into environment variables for household flexible loads, electric heating and EV battery charging and discharging. Based on the environment step, the next state s_i^{t+1} and reward r^t are obtained. Each agent i also estimates a critic network that estimates the individual action-value function Q_i using a linear network with one hidden layer. The Q_i values are then combined into the centralised action-value function Q_{tot} via a non-linear monotonic mixing function approximated by a mixing network with non-negative weights as:

$$Q_{\text{tot}}^{\pi}(\boldsymbol{\tau}, \boldsymbol{a}, s; \boldsymbol{\phi}, \psi) = g_{\psi}(s, \{Q_i^{\pi_i}(\tau_i, a_i; \phi_i)\}_{i=1}^{n}) \tag{5.1}$$

where $\boldsymbol{\phi}$ and ϕ_i are parameters of the centralised action-value function Q_{tot}^{π} and agent-wise utilities $Q_i^{\pi_i}$, respectively, and g_{ψ} is a non-linear monotonic function parameterised as a mixing network with parameters ψ, as in QMIX [217]. To evaluate the policy, the centralised but factored critic is trained by minimising the following loss:

$$\mathcal{L}(\boldsymbol{\phi}, \psi) = \mathbb{E}_{\Theta}[(y^{\text{tot}} - Q_{\text{tot}}^{\pi}(\boldsymbol{\tau}, \boldsymbol{a}, s; \boldsymbol{\phi}, \psi)^2] \tag{5.2}$$

where $y^{\text{tot}} = r + \gamma Q_{\text{tot}}^{\pi}(\boldsymbol{\tau'}, \boldsymbol{\pi(\tau'; \theta^-)}, s'; \boldsymbol{\phi^-}, \psi^-)$, Θ is the replay buffer, and $\boldsymbol{\theta^-}$, $\boldsymbol{\phi^-}$ and ψ^- are the parameters of the target actors, critic, and mixing network, respectively.

To update the decentralised policy of each agent, a centralised policy gradient estimator is used to optimise over the entire joint action space:

$$\nabla_{\theta} J(\boldsymbol{\pi}) = \mathbb{E}_{\Theta}[\nabla_{\theta} \boldsymbol{\pi} \nabla_{\boldsymbol{\pi}} Q_{\text{tot}}^{\boldsymbol{\pi}}(\boldsymbol{\tau}, \pi_1(\tau_1), ..., \pi_n(\tau_n), s)] \tag{5.3}$$

where $\boldsymbol{\pi} = \{\pi_1(\tau_1;\theta_1), ..., \pi_n(\tau_n;\theta_n)\}$ is the set of all agents' current policies, and all agents share the same actor network parameterised by θ.

While back-propagation via the critic network is performed to update the network weights during training, only a feed-forward flow through the actor network from observation to action is needed during execution.

5.2.2.1 Convolutional networks

A convolutional layer allows for enhanced feature extraction within actor networks when learning the selection of action time series for the day ahead.

Initially developed for image data analysis by Lecun et al. [273], Convolutional Neural Networks (CNNs) gained prominence for their remarkable ability to acquire intricate feature representations. This capability is achieved by applying a set of filters to the input image, allowing them to extract meaningful features. The discrete formulation of the convolution is [274]:

$$(f * g)[n] = \sum_{m=-\infty}^{\infty} f[m] \cdot g[n-m] \tag{5.4}$$

where $(f * g)[n]$ is the result of the 1D convolution operation at position n, $f[k]$ is the value of the filter or kernel at position m, and $g[n-m]$ is the value of the input signal at position $n-m$.

While CNNs were initially tailored for image analysis, their application to time series data has been emerging. Several examples in the literature advocate the use of CNNs for time series classification [156, 275, 276] and generation [277].

The extensive success of CNNs in RL, exemplified by their performance in playing Atari games as demonstrated by Mnih et al. [278], where they effectively extracted environmental information from image screenshots, underscores their versatility. Given their potential effectiveness in discerning patterns within time series data, CNNs offer exciting prospects in the realm of deep RL [279]. This potential is particularly relevant when planning actions across time series.

5.2.2.2 Supervised loss

Inspired by [280], this thesis proposes a mixed optimisation-informed centralised but factored critic approach that incorporates a supervised loss so agents can learn from both demonstrator and exploration data. The supervised loss enables the agent to learn to mimic the expert demonstrations (here, the convex optimisation results), while the temporal difference loss enables the agent to learn from its own experience generated through directly interacting with the environment. This mechanism thus combines "expensive" demonstration data from convex optimisations and "cheap" exploration data from the simulated environment to guide the learning of multi-agent cooperation. A weighted penalty is added to the loss for actions that deviate from the demonstrator data:

$$\delta_{\textbf{supervised loss}} = C \sum_{t} \left(a^{t}_{i,\textbf{demonstrator}} - a^{t}_{i,\textbf{exploration}} \right)^2 \tag{5.5}$$

This work investigates whether this could guide the agents' learning and improve coordination performance, as the exploration space and coordination challenges are particularly potent in MARL.

5.2.2.3 Exploration

For the first n_{uniform} steps, the actions are sampled from a uniform-random distribution. After this initial phase, Gaussian noise is added to the actions to promote exploration, with standard deviation σ_{action}.

Moreover, the network weights are initialised by drawing from a normal distribution with variance $\sigma^2_{\text{network}}$.

5.2.2.4 Observation space

The observation space is continuous and consists of the grid cost coefficients for 24 hours ahead $o^t_i = \{\mathrm{C}^t_{\mathrm{g}}, ...\mathrm{C}^{t+N}_{\mathrm{g}}\}$.

5.2.2.5 Action space

The action space is continuous and consists of the three actions defined in Section 3.2.2 for the 24 hours ahead.

$$a_i^t = \begin{bmatrix} a_{\text{bat},i}^t & \dots & a_{\text{bat},i}^{t+N} \\ a_{\text{heat},i}^t & \dots & a_{\text{heat},i}^{t+N} \\ a_{\text{cons},i}^t & \dots & a_{\text{cons},i}^{t+N} \end{bmatrix} \tag{5.6}$$

5.2.2.6 Reward

This chapter defines the global reward as the negative sum of grid, distribution and battery costs, defined in Section 3.1.2.

$$r^t = -\left(c_{\text{g}}^t + c_{\text{d}}^t + c_{\text{b}}^t\right) \tag{5.7}$$

5.2.3 Hyper-parameter tuning

The same methodology is used as in Section 4.2.3 to tune the hyper-parameters. It is worth noting that hyper-parameters are more numerous in deep neural networks relative to tabular methodologies, with consequently more time and resources required to optimise the parameters.

5.3 Experiments

This section first presents the main results and insights offered by the parameter tuning process Section 5.3.1. Then, the results of experiments simulating the coordination of residential energy flexibility using FACMAC are presented and analysed. Namely, the coordination performance and computational scalability of the centralised but factored critic framework are compared to those of optimisation-informed learning in light of their structural differences in Section 5.3.2, and their relative reliabilities are assessed in Section 5.3.3. Following these results, Section 5.3.4 more widely discusses the advantages and disadvantages of each approach and how to select between them for the coordination of DERs in this thesis.

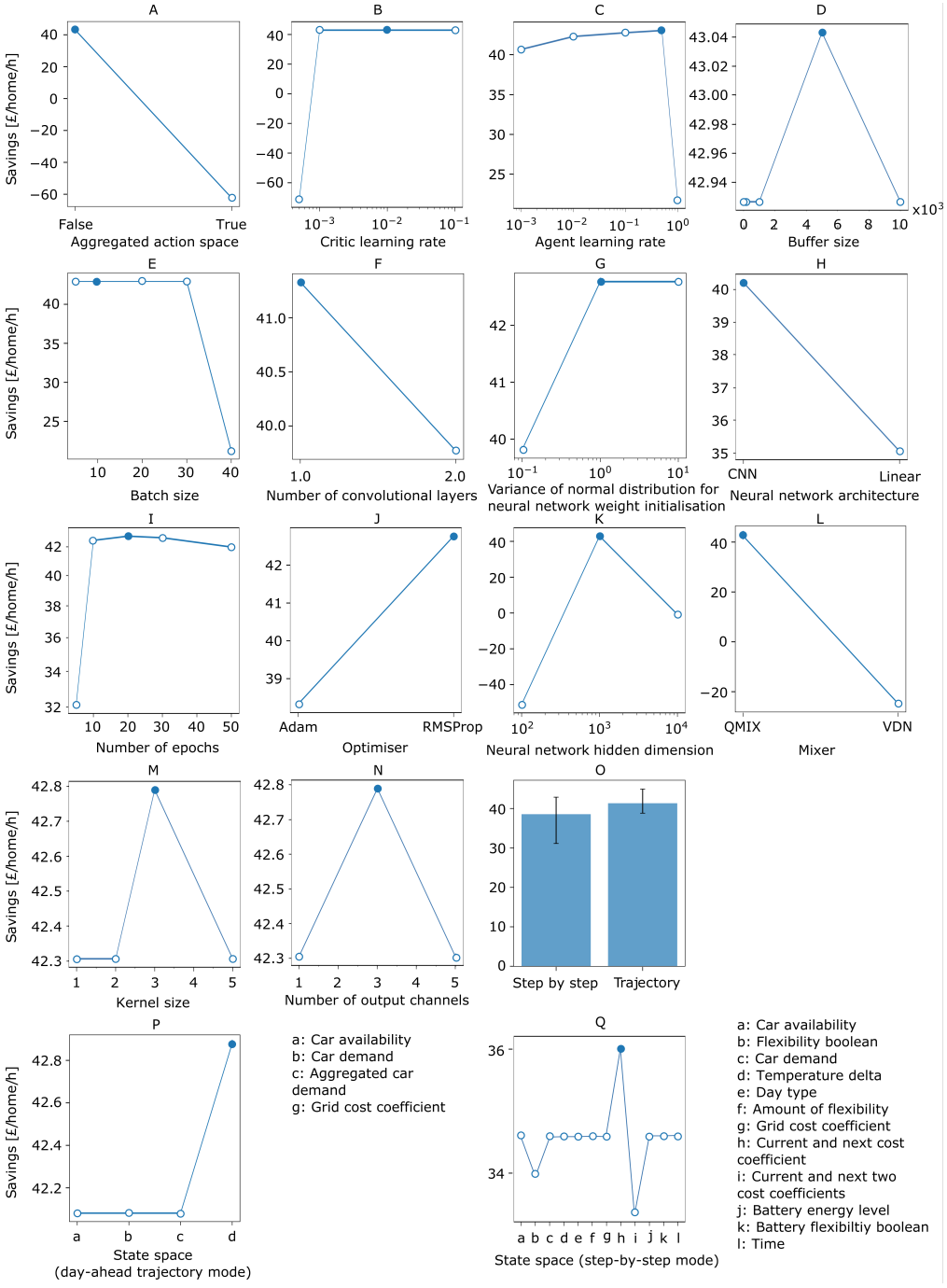


Figure 5.2: Parameter tuning using deep learning with a centralised but factored critic. Savings obtained per home and month, only varying one hyper-parameter at at a time. The highest savings are indicated by the full dot. All sensitivities use the trajectory mode (where actions and states are defined for the whole 24-hour period ahead as opposed to one time step at a time) unless specified otherwise, as this was the best-performing mode. When comparing the two modes (sub-figure O, P, Q), the hyper-parameters for each option have previously been optimised

5.3.1 Parameter tuning

The main results of the sensitivity analyses are tabulated in Table 5.1 and illustrated in Figure 5.2. Six takeaways from this analysis are:

1. Similarly to in Chapter 4, a disaggregated action space provides superior performance, with three action variables for household consumption, heating and EV charging $a^t_{\text{cons},i}$, $a^t_{\text{heat},i}$ and $a^t_{\text{bat},i}$ (Figure 5.2-A).

2. Choosing actions for the whole day ahead based on day-ahead grid prices provides better performance than step-by-step decision-making (Figure 5.2-O).

3. The optimal agent learning rate (Figure 5.2-C) is an order of magnitude larger than the critic learning rate (Figure 5.2-B). A lower learning rate for the critic ensures that its updates are more stable and less prone to drastic changes, which can lead to a smoother learning process. Moreover, it helps mitigate the risk of overestimating the temporal difference (TD) error, which can lead to convergence issues. The critic can provide a more consistent assessment of the agent's actions and help prevent over-fitting to noisy or irrelevant information in the learning process.

4. Increasing the variance of the normal distribution for neural network weight initialisation led to improved performance (Figure 5.2-G), by encouraging the network to explore a broader range of weight values at the beginning of training. This exploration helps avoid convergence to suboptimal solutions, as the network has a higher chance of finding better regions of the weight space. It also helps address vanishing or exploding gradients, a common problem in deep networks where gradients become too small or too large during back-propagation, hindering the effective update of weights and training of the network. Higher weight variances can help mitigate these issues by providing more "room" for gradients to propagate through the layers without being diminished or amplified too quickly.

5. Using the QMIX factorisation was key in achieving coordination (Figure 5.2-L). Simply adding up individual value functions using VDN could not achieve savings relative to the baseline. As a non-linear extension of VDN, QMIX can represent more extra state information during training and a much richer class of action-value functions, as discussed in Section 2.2.3.2.

6. Convolutional neural networks provided superior coordination performance relative to linear ones (Figure 5.2-L). CNNs are designed to capture spatial patterns in data, making them well-suited for tasks where relationships between inputs, such as sequential electricity prices, play an important role. In sequential decision-making tasks, a linear model might struggle to capture these relationships effectively. In addition, this pattern recognition is designed to be invariant to translations. Even if specific home parameters or the exact time of day might vary, the CNN can learn to focus on relevant features without being affected. This is because CNNs automatically learn hierarchical representations of the input data at different levels of abstraction. This feature learning capability allows the CNN to identify meaningful patterns in the sequential data (e.g. prices and energy demand time series) to make decisions over multiple time steps.

Description	**Value**
Depreciation rate	$\gamma = 0.85$
Critic learning rate	$\alpha_\phi = 1 \times 10^{-2}$
Actor learning rate	$\alpha_\theta = 5 \times 10^{-1}$
Buffer size	5×10^3
Neural network optimisation algorithm	RMSprop [281]
Batch size	10
Variance of normal distribution for network weight initialisation	$\sigma^2_{\text{network}} = 1$
Convolutional layer kernel size	3

Table 5.1: Factored but centralised critic learning parameters

5.3.2 Results

Learning experiments are performed over 20 epochs consisting of an exploration, an update and an evaluation phase, as in Section 4.3.1, to maintain comparability.

First, the environment is explored over two training episodes of duration $|\mathcal{T}| = 24$ hours. Exploration data is generated by HEDGE, and optimisations are performed on this input data in the case of optimisation-learning. Then, neural network weights are updated based on the rules presented in Section 5.2.2. Finally, new testing data is generated and an evaluation is performed. Ten repetitions are performed such that the learning may be assessed over different trajectories.

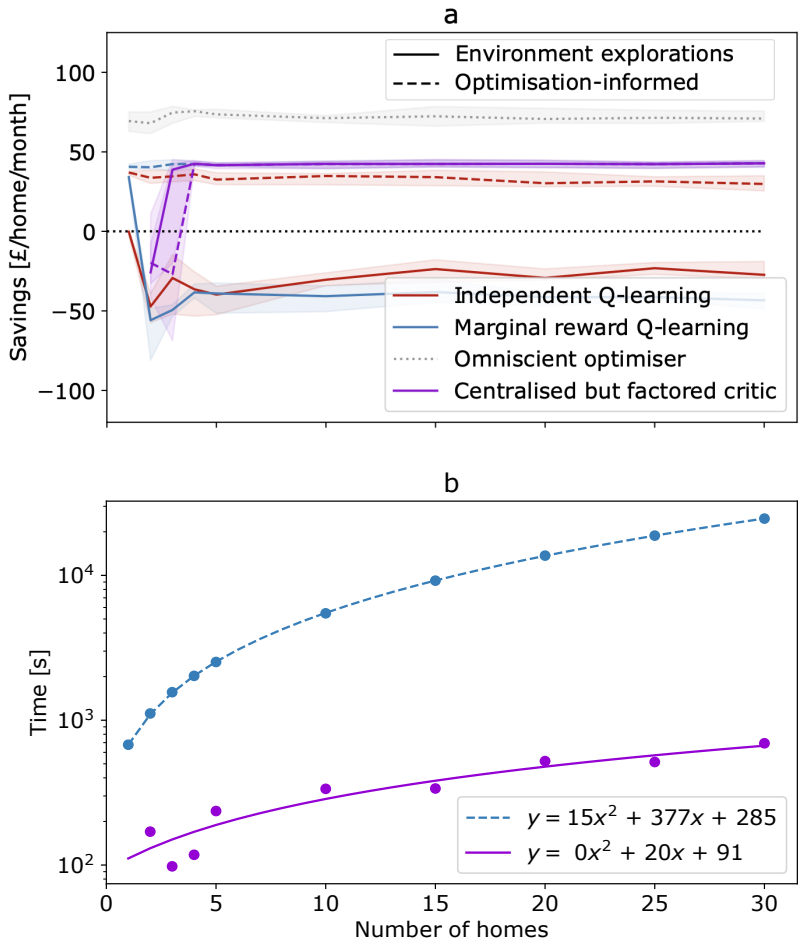


Figure 5.3: (a) 50th percentile savings per home and month relative to the baseline scenario as the number of agents increases over ten repeats. Full lines correspond to training using random environment explorations only. Dotted lines use experience from convex optimisations for training. The shaded areas show the spread between the 25th and 75th percentile values over the ten repeats. (b) Corresponding required training time for the ten repeats for the two methodologies that achieve coordination at scale, namely optimisation-informed independent Q-learning and the centralised but factored critic approach. Functions are fitted using SciPy's optimise.curve_fit non-linear least squares function [270].

Figure 5.3(a) shows the total system savings achieved per home and month as the number of agents trained increases. These savings are computed relative to a baseline where all agents are inflexible, with EVs charged immediately and no flexible loads delayed.

The centralised but factored critic methodology (purple lines) performs worse than the baseline (negative savings) when the number of agents is lower than four due to the increased variability exhibited by an insufficient number of agents, which poses challenges in estimating a reliable mixing network and learning robust policies. The performance then increases at scale, demonstrating comparable coordination performance to the optimisation-informed independent learning approach with marginal rewards (blue dotted line), whilst overcoming its dependency on optimisations and additional marginal rewards computations.

Figure 5.3 thus shows that optimisation-informed FACMAC (incorporating the supervised loss, purple dotted line) does not achieve superior performance relative to FACMAC. This demonstrates that the centralised but factored critic alone already provides an equivalent coordination mechanism for the cooperation of the agents in the system, and the use of optimisation data adds no additional value. This suggests that FACMAC already learns equivalent coordination, but with widely different computational costs relative to optimisation-informed IQL, due to their differing architectures.

On the one hand, in the interior point method, which is used to solve non-linear continuous problems in convex optimisation solvers, the inverse of the Jacobian of the Karush–Kuhn–Tucker (KKT) conditions must be computed in each Newton-Raphson update step. As this includes derivatives with respect to all decision variables of the problem for all constraints and the objective function [282], the Jacobian grows with $O(n^2)$. This step therefore bears a high computational burden, which can be prohibitive as the system size grows [283], particularly in machine learning applications where numerous optimisations must be performed to generate training data.

On the other hand, the centralised but factored critic can capture a global understanding of the impact of individual policies on the system via the centralised action-value function, using extra information (e.g. global state and joint action) available only during training. In the residential DER coordination problem, there is particularly high value in cooperation signals that take a global view of the system, as the global rewards are highly dependent on the cumulative impact of multiple agents taking actions simultaneously. The backward propagation of gradients from the centralised but factored critic to the individual actor networks can then guide the optimisation of individual policies in the absence of full state information at the scale of individual agents [223]. FACMAC can also enable more scalable learning as the centralised action-value function is represented as a combination of individual action-value functions, which condition on much smaller local observations and actions. As the number of value networks only grows linearly with the number of homes, the learning has computational complexity growing linearly with the number of homes $O(n)$. The factored critic uses information more efficiently as network weight updates use state space information that has the most impact on the global value of actions taken by agents. The updates structurally take into account the partial observability of agents taking actions, and aim to update only the knowledge necessary to know which actions agents should take and when. The back-propagation from the centralised but factored critic thus emulates the global optimisation, given the partial observability of the agents taking actions: each back-propagation includes a Jacobian-gradient product of the value error with respect to the weights of the networks for each operation in the graph.

The optimisation-informed independent learning and the centralised but factored critic approaches thus mirror each other in their provision of a global coordination mechanism and achieve very similar performance, while having structural differences which lead to varying computational efficiencies. In both cases, the Jacobian plays a role, and both marginal rewards and the backward propagation from the centralised but factored critic aim to improve learnability, sending personalised rewards to each agent that best represent their contribution to the global reward.

Overall, the total costs could thus be reduced by £42.42 per home and month on average for 30 homes by the centralised but centralised critic architecture. This represents a 20.0% reduction from the baseline, i.e. 57.3% of the upper bound savings achieved by an omniscient and time-travelling optimiser. FACMAC did not achieve satisfactory coordination below four agents: it was designed for scale, when more individual critic networks can yield better centralised critic factorisation. Figure 5.3(b) shows the optimisation-informed approach required second-order polynomial computational time as the number of agents increases, whereas the centralised but factored critic methodology required approximately first-order polynomial time $O(n)$. For 30 homes, the factored critic approach thus required 34.0 times less computational time. As shown in Figure 5.4, this superior computational scalability allows the FACMAC methodology to be used at the feeder level (100 homes), as was the aim of this thesis.

5.3.3 Reliability

Figure 5.5 compares the reliability of the independent learning, optimisation-informed and centralised but factored critic algorithms in an experiment with 30 homes, using the metrics presented in Section 4.3.5.

It is found that FACMAC has lower Long-term Risk across Time (LRT) and Dispersion across Time (DT), and equivalent savings, Dispersion across Runs (DR), Risk across Runs (RR) and Short-term Risk across Time (SRT) than IQL. Therefore, it not only provides improved savings but also reduces risk and variability.

5.3.4 Optimisation-informed tabular independent learning vs. deep MARL with factored but centralised critic

The utilisation of deep reinforcement learning techniques employing large neural networks may not be an optimal approach to solving relatively simple problems, which is why tabular methods retain their utility in such scenarios. Particularly, if the number of state and action combinations is sufficiently small to fit in memory and be visited multiple times within a reasonable duration, then tabular methods

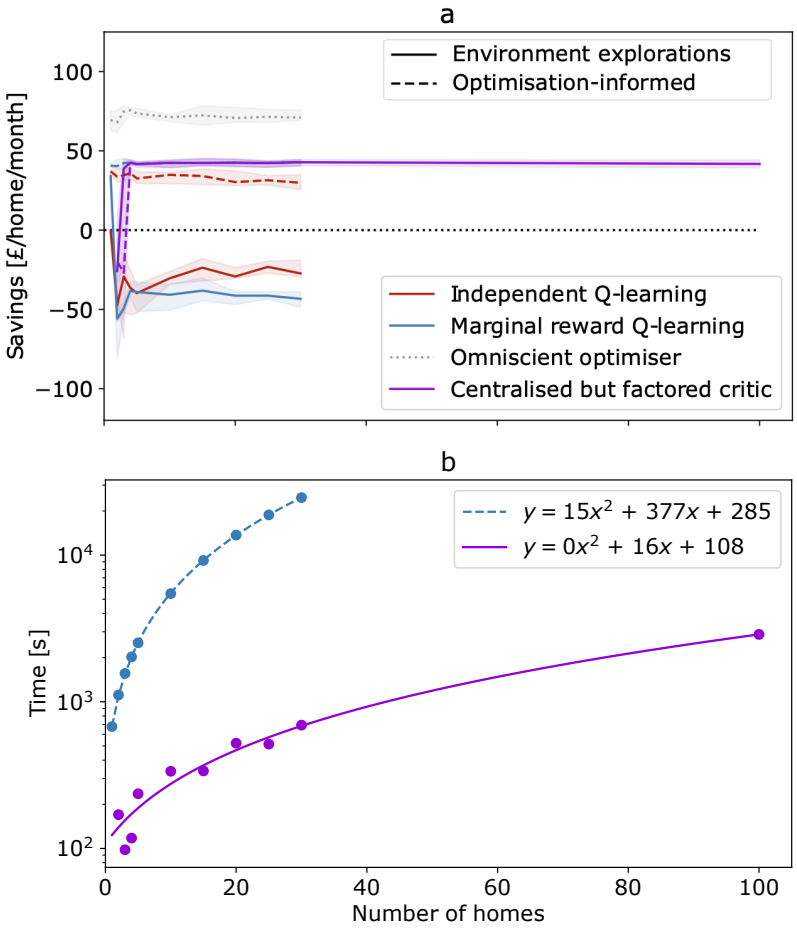


Figure 5.4: (a) 50th percentile savings per home and month relative to the baseline scenario as the number of agents increases over ten repeats. Full lines correspond to training using random environment explorations only. Dotted lines use experience from convex optimisations for training. The shaded areas show the spread between the 25th and 75th percentile values over the ten repeats. (b) Corresponding required training time for the ten repeats for the two methodologies which achieve coordination at scale, namely optimisation-informed independent Q-learning and the centralised but factored critic approach. Extended to coordinating 100 homes in the case of FACMAC.

can offer convergence guarantees that are not achievable by approximate methods. Consequently, tabular approaches are generally favoured when they are applicable.

Moreover, tabular methods have fewer parameters compared to deep learning models. As a result, their development is generally less complicated as there is no need for extensive parameter tuning, which often plays a critical role in determining the performance of deep learning models.

However, as problems become more complex and realistic, they no longer fit

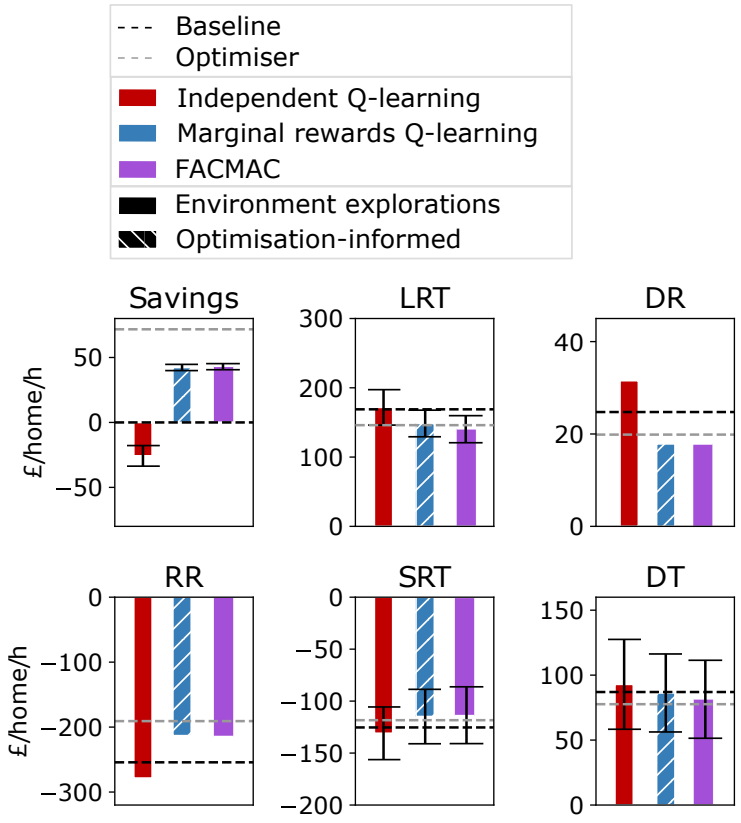


Figure 5.5: Reliability metrics: independent learning vs. FACMAC in experiments with 30 homes.

the tabular approach. Gaining a thorough understanding of tabular value-based methods is still valuable since they serve as the fundamental building blocks for more intricate deep-learning methods. To some extent, these methods represent optimal solutions that deep reinforcement learning aims to approximate, and the development of tabular solutions can inspire modifications and adaptations to neural network-based approaches. An analysis of the use of tabular independent learners as the system size grows has thus highlighted the need for a coordination mechanism. While this coordination mechanism could be provided by the use of optimised training data within the independent learner tabular framework, this is computationally expensive at scale.

On the other hand, the centralised but factored critic approach was designed with a primary emphasis on scalability. While it achieves poor performance for fewer than four agents, more agent networks can be combined into more reliable estimates

of the global critic efficiently. This can be explained by the fact that the variability for a small number of homes is high. These results are consistent with previous work, which showed that both the after-diversity maximum demand and the variance of residential power consumption profiles are high for fewer than four aggregated homes, and drop steeply to stabilise for higher numbers [245]. As demonstrated by the central limit theorem, as the number of diverse homes increases, the spread is inversely proportional to the number of homes and more robust policies can be learned. The adapted FACMAC theorem can achieve equivalent performance at scale with much lower computational requirements, and be more data efficient thanks to the use of batch RL techniques that store and reuse past interactions. Not only does the factored but centralised critic approach require shorter computational time, it also grows at first-order $O(n)$ in lieu of second-order $O(n^2)$ polynomial time as the system size increases.

This thesis therefore recommends using the optimisation-informed IQL algorithm for systems with fewer than four homes and the centralised but factored critic approach beyond. This is the approach that will adopted to assess the challenges and benefits of implementation at scale of these algorithms in Chapter 6.

5.4 Concluding remarks

This chapter has employed a cooperative multi-agent actor-critic framework that learns decentralised policies with a centralised but factored critic, to address the pitfalls of both the poor coordination performance of independent learners, and the intractability at scale of optimisations or centralised critic estimation.

This approach allows the coordination to reach the secondary substation level, as was aimed at in this thesis (Table 5.2). Compared with the previous independent learning approach, which learns from global optimisation results, the centralised but factored critic yields inferior coordination for fewer than four agents, but similar coordination at scale. Its computational time requirements are 34 times lower for 30 homes, and growing only at first-order rather than second-order polynomial time.

This improved scalability opens the way for coordinating flexible energy resources such as electric vehicles and heating in a fully distributed manner.

Chapter 6 now analyses in more detail the potential challenges and impacts of implementing such an approach on electricity networks and energy users.

Criterion addressed	**Verification**
3. Computational scalability: A successful coordination will be considered to have overcome the scalability challenges presented in Section 1.2.3, if it only imposes minimal and constant distributed computation burden during implementation as the system size increases up to at least the feeder level (∼100 homes), and with only first-order computational requirement growth for training $O(n)$.	✓

Table 5.2: Current assessment of algorithm success criteria set in Section 1.3

6

Challenges and benefits of implementation at scale

Contents

Having successfully developed a robust multi-agent coordination algorithm suitable for the decentralised coordination at scale of residential energy without sharing private data, the following research question is investigated: *Can the algorithms achieve positive impacts for energy users and the grid?* This chapter addresses crucial aspects of the deployment of the algorithm within real-world contexts.

Firstly, Section 6.1 compares the impact on voltage constraints of uncoordinated

DERs, grid-unaware coordinated DERs[1], and DERs cooperatively managing voltage constraints. The potential benefits of adopting cooperative grid management mechanisms to alleviate pressure on the grid infrastructure are analysed in a case study in the simulated IEEE European Low Voltage Test Feeder [265].

Subsequently, Section 6.2 evaluates the repercussions of the energy flexibility coordination system on energy users. The cost savings achieved by participants in scenarios with and without cooperative voltage management are analysed. Understanding the implications for end-users is crucial for gauging the acceptability and efficacy of the proposed coordination approach.

Finally, the robustness and generalisability of the positive impacts of MARL-based implicit cooperation under environment variations are assessed in Section 6.3. Specifically, we consider cases when the number of homes, both coordinated and uncoordinated, differs from the base case, as well as a scenario where policies are implemented in different years, with varying energy prices and weather conditions. Evaluating policies in diverse scenarios helps enhance their practical utility and trustworthiness.

6.1 Assessing and managing network impacts

This section first assesses the impact of grid-unaware DER coordination on the network in Section 6.1.1. Then, it explores the potential of MARL-based cooperative grid management in Section 6.1.2. PandaPower is used to simulate the impact of the actions taken by RL agents. This open-source Python tool offers a Newton-Raphson power flow solver, accelerated with just-in-time compilation [242].

6.1.1 The impact of grid-unaware DERs on the network

DER coordination will occur within physical distribution networks, often designed and constructed before the advent of the types of DERs coordinated in this thesis.

[1]DER coordination without penalties for voltage constraint violations in the reward or objective function.

We consider a model of such a real-life distribution network, the IEEE European Test Feeder, which contains 55 homes and is depicted in Figure 6.1.

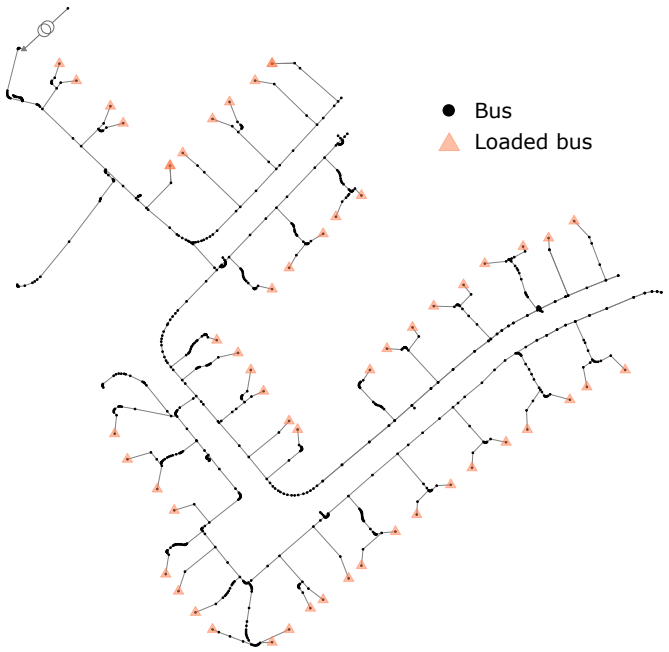


Figure 6.1: IEEE Low Voltage Test Feeder. Loaded buses on the network correspond to the homes in the experiments.

While ownership of DERs leads to cost savings for energy users (as evidenced in Chapters 4 and 5), they can also exacerbate the frequency of voltage constraint violations. Figure 6.2 illustrates that, without coordination, existing networks are more susceptible to voltage constraint violations. This, in turn, heightens the necessity for expensive voltage infrastructure upgrades and increases the frequency of operational disruptions. Even without explicitly considering grid constraints, the MARL-based coordination of DERs developed in Chapters 4 and 5 diminishes the likelihood of voltage constraint violations compared to this baseline scenario. This reduction results from the maximisation of the other reward components, which tends to "smooth out" import and export profiles, as exemplified in Section 4.3.4. Nevertheless, some voltage constraint violation increases persist compared to scenarios with no DER adoption.

When examining the three types of DERs modelled in this thesis (electric heating, EVs, and flexible loads) individually, Figure 6.3 reveals that coordinating

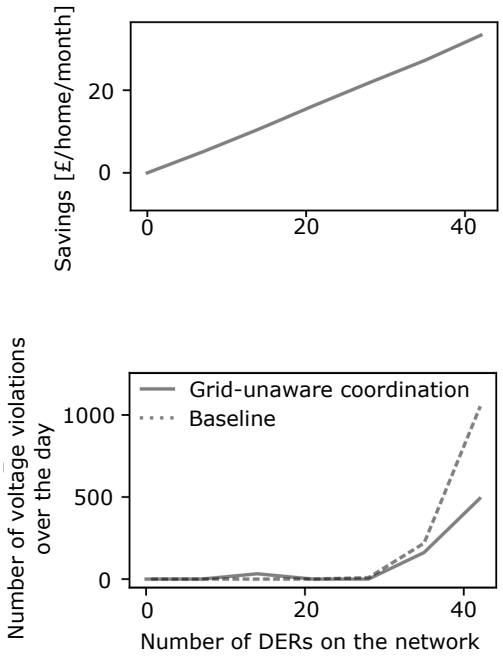


Figure 6.2: Coordination savings and number of voltage constraint violations throughout a simulated day, as the number of nodes possessing DERs (heat pumps, electric cars and flexible loads) increases.

DERs without explicitly accounting for voltage constraints reduces the occurrence of voltage constraint violations for heat pumps and household loads, and increases it in the case of EVs. This increase is due to the objective function's emphasis on minimising energy costs and promoting energy arbitrage. This promotes the practice of minimising car battery charging during higher price intervals and charging it when prices are lower, resulting in uneven charge and discharge profiles.

Furthermore, it is noteworthy that while 2 MWh/day of household loads (equivalent to 200 homes) can be integrated into the test network without incurring voltage constraint violations, challenges arise from 1 MWh/day in the case of EVs (equivalent to 140 homes with EVs) and heat pumps (50 homes). This discrepancy highlights that the number of network constraint violations is not directly proportional to the total daily load. Instead, it is influenced by various factors, including the shape of the load profile, the potential for flexibility within the loads, and the degree of correlation among loads within a neighbourhood.

Figure 6.4 demonstrates that MARL-based coordination results in a decrease

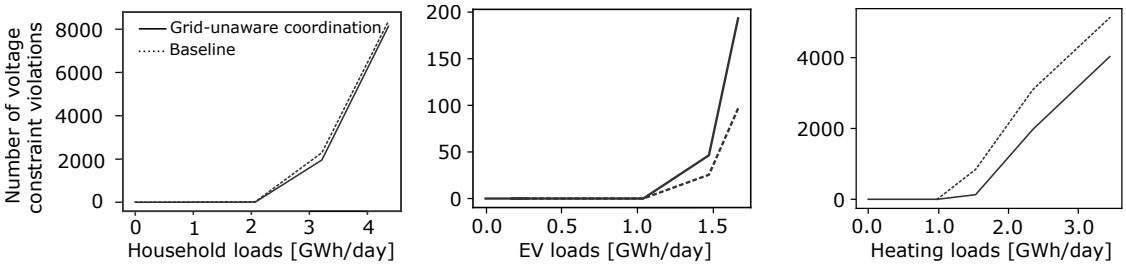


Figure 6.3: Coordination savings and number of voltage constraint violations throughout a simulated day, as the number of nodes possessing DERs (heat pumps, electric cars and flexible loads) increases.

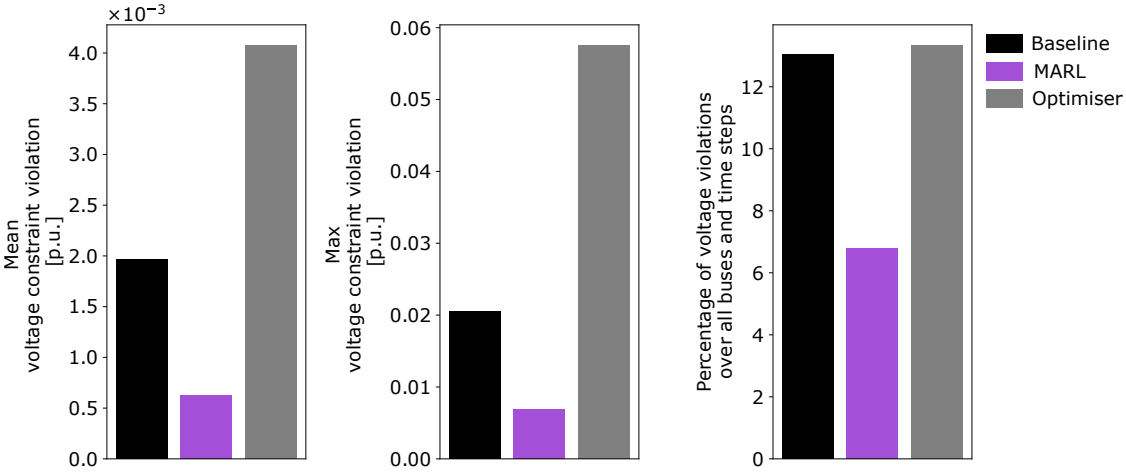


Figure 6.4: Voltage constraint violation metrics for the inflexible baseline, MARL strategy and central optimiser that do not consider voltage deviations in their objective functions and rewards.

in the mean, maximum, and overall number of voltage constraint violations, even without a specific objective related to network constraint management. Additionally, the plot clearly shows that the central optimiser, which also does not incorporate voltage deviations in its objective function, substantially increases the occurrence of such violations. This phenomenon can be ascribed to its emphasis on prioritising energy arbitrage opportunities, which in turn leads to fluctuations in the network loading. Figure 6.5 illustrates how these variations in voltage constraint violation occurrence manifest throughout the day.

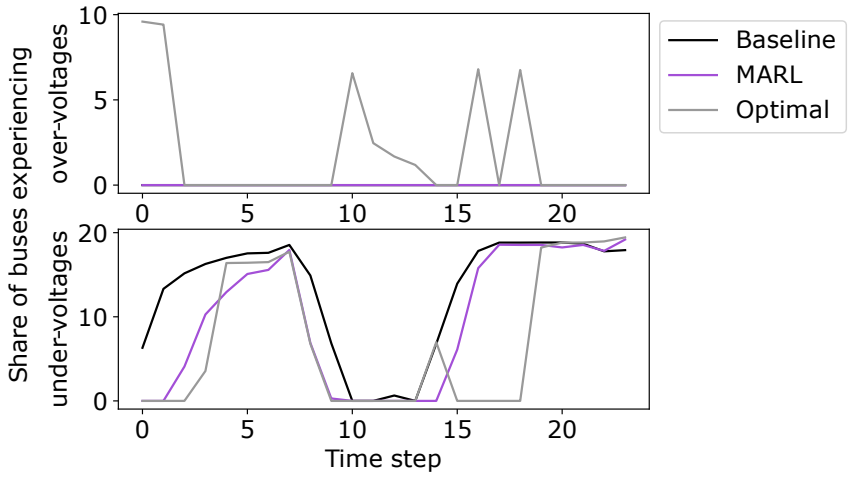


Figure 6.5: Voltage constraint violations over the day using the methodology presented in Chapter 5 with no voltage management for 55 homes in the IEEE Low Voltage test feeder.

6.1.2 Grid-aware coordination

This section examines the potential of the MARL-based decentralised voltage constraint management in distribution networks. As elucidated in Section 6.1.1, MARL-based coordination can pose challenges to the efficient and secure operation of distribution networks, leading to potential voltage constraint violations and line congestions [69, 102]. On the other hand, if appropriately managed, DERs have the potential to offer valuable services to system operators [72]. This section demonstrates that the outcomes of MARL-based cooperative voltage constraint management are improved by continuously monitoring real-time voltage levels and sharing this information with DER agents.

In this analysis, the reactive power control RL action presented in Section 3.1.5 is enabled for EVs. Utilising reactive power from DERs can help mitigate the voltage elevation resulting from real power injections [284]. Moreover, a voltage penalty c_{v}^{t} proportional to voltage constraint violations is added to the reward function (see Section 3.1.2). Given the statistical RL approach adopted in this thesis and the frequent occurrence of under-voltages in the baseline scenario, the minimum voltage limit is set as $\underline{\mathrm{v}} = 0.96$ [p.u.]. This setting allows voltage values to start being penalised even before the usual lower limit of -6% from the

nominal value, to promote the learning of behaviours that limit the likelihood of larger voltage constraint violations.The total system costs to minimise include grid energy (c_{g}^{t}, Equation (3.2)), battery (c_{b}^{t}, Equation (3.5)), and voltage (c_{v}^{t}, Equation (3.6)) costs [£]:

$$\max F = \sum_{\forall t \in \mathcal{T}} \hat{F}_t = \sum_{\forall t \in \mathcal{T}} - \left(c_{\mathrm{g}}^{t} + c_{\mathrm{b}}^{t} + c_{\mathrm{v}}^{t} \right). \tag{6.1}$$

Figure 6.6 illustrates the relationship between the share of all buses and time steps experiencing voltage constraint violations and the voltage constraint violation penalty coefficients $\mathrm{C}_{\overline{\mathrm{v}}} = \mathrm{C}_{\underline{\mathrm{v}}}$. These values are compared to a baseline scenario where all agents remain passive, i.e., do not use their flexibility, with EVs charged immediately and no flexible loads delayed.

Two experimental set-ups are compared. In the first, the agents' state spaces include only the grid price coefficient C_{g}^{t}, as in Chapters 4 and 5. In the second, the state space is extended to include the minimum voltage value across all network buses at the start of each time step [2]. The latter addition is shown to be crucial in enabling effective voltage constraint management[3]. The results indicate a reduction in the number of voltage constraint violations over the day by 48.8% as the voltage constraint violation penalty coefficients $\mathrm{C}_{\overline{\mathrm{v}}}$ and $\mathrm{C}_{\underline{\mathrm{v}}}$ increase from $1{\times}10^{-5}$ to $1{\times}10^{2}$ [£/p.u.], to reach a 43.1% reduction in the number of violations relative to the baseline.

Figure 6.7 thus illustrates that, relative to the baseline, the mean voltage constraint violation, maximum violation, and the total number of violations could be reduced by 63.5%, 63.6%, and 43.1%, respectively.

Figure 6.8 depicts how this reduction in voltage constraint violation occurrence is achieved throughout the day for both optimisation- and MARL-based control.

[2]While traditional network analysis often revolves around monitoring voltage levels at specific bus locations, this approach can be myopic, as it may not capture the broader implications of network disturbances. In contrast, the lowest voltage value serves as an early common indicator of the network's overall stability, as voltage deviations propagate and accumulate down the interconnected system. This can help agents respond promptly to overall network disturbances and implement targeted voltage regulation and control strategies.

[3]It is worth noting that, to provide real-time voltage values to the state of agents, this experiment was conducted in a step-by-step RL implementation, as opposed to a daily trajectory as in Chapter 5.

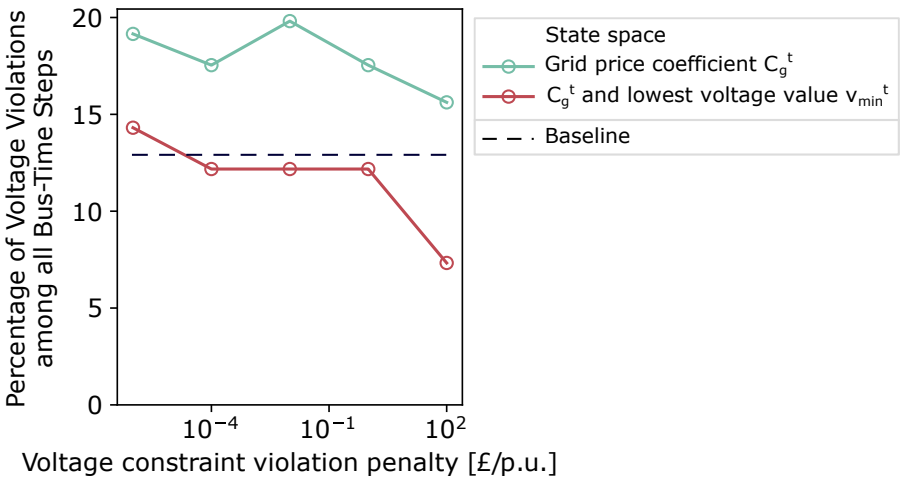


Figure 6.6: Number of voltage constraint violations with and without voltage information accessible to cooperating households.

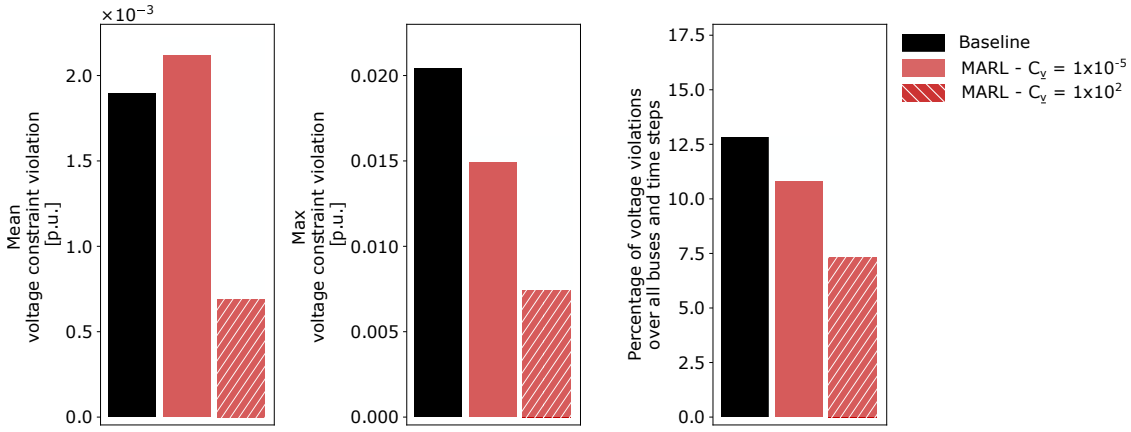


Figure 6.7: Voltage constraint violation metrics as the voltage constraint violation penalty is increased relative to the baseline.

The figure reveals that, while the optimiser has no under-voltages at 4 pm, when the under-voltages of the MARL-based controller peak, other peaks occur earlier at 3 pm and later at 9 pm. Furthermore, the optimiser causes several swings throughout the day, which may be challenging to predict and manage. This behaviour is attributed to the optimiser's continuous trade-offs in the optimisation objectives not only in voltage constraint management but also in energy arbitrage opportunities.

In contrast, the MARL strategies maintained the overall baseline shape, though with reduced over-voltages. Notably, contrary to the optimal and MARL-based scenarios with no voltage management, no over-voltages occurred. The MARL-

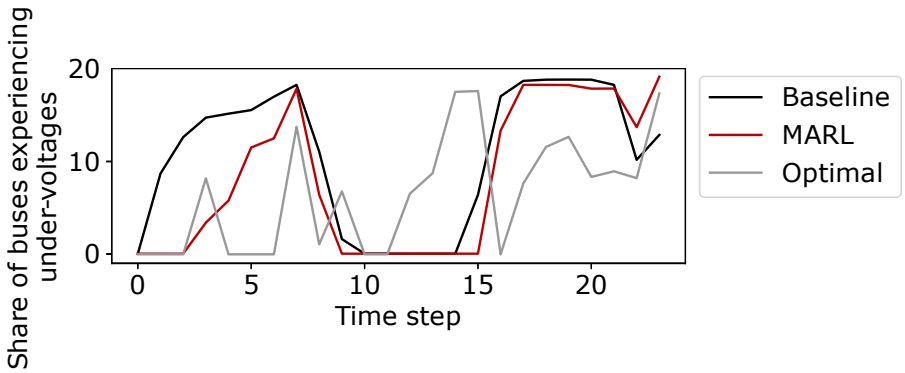


Figure 6.8: Number of voltage constraint violations over the day for MARL-based coordination with cooperative voltage constraint management. The minimum voltage information is shared with agents, and voltage penalty coefficients of 1×10^2 are used. Comparative profiles are provided by an inflexible baseline and a centralised, omniscient convex optimisation.

based coordination shaves both morning and evening peaks and yields 63.6% of the voltage constraint violation occurrence reduction achievable by the optimiser.

The resulting trade-offs in other private costs for energy users participating in MARL-based cooperative voltage constraint management are presented in Section 6.2.

6.2 Value for energy users

As argued in Chapter 1, it is imperative to assess whether MARL algorithms that could be deployed in individual households entail an increased risk to energy users. While energy users have the potential to gain from cooperating, both from a community and an individual perspective [73], maximising overall system utility may sometimes come at the expense of individual energy users [76].

6.2.1 Without voltage management

The primary objective of active participants in local energy markets is to reduce their individual costs while maintaining their comfort or utility. Costs to minimise include energy bill payments and asset degradation (managing battery life).

While the statistical approach proposed in this thesis does not provide theoretical guarantees on the minimum savings achieved by energy users, empirical evidence in

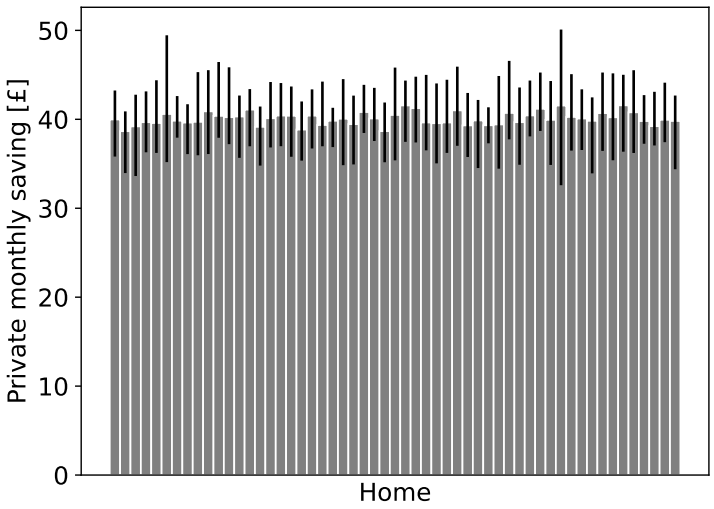


Figure 6.9: Private savings in energy bills and battery depreciation cost relative to the baseline over one month without cooperative voltage management for 30 participants. The bar is the average value, and the black vertical lines are the spread from the minimum to the maximum monthly savings over the ten repeats.

Figure 6.9 demonstrates that all prosumers consistently achieve savings in private energy bill and battery depreciation costs, even in the worst-case scenarios across ten repeats. This figure corresponds to the case without cooperative voltage management, as in Chapter 5.

All homes achieve significant savings of £40.08, or 20.9% of their baseline monthly private costs. Moreover, these savings can be deemed *fair*, as the homogeneous set of homes all obtain similar savings, with merely a £2.90 difference between the minimum and maximum average monthly savings per home. Thus, even in the absence of additional remuneration or value-sharing mechanisms, all prosumers readily benefit from participating in MARL-based cooperation without voltage management.

6.2.2 With voltage management

A similar analysis is provided for the case with cooperative voltage management and a penalty coefficient of 1×10^2 [£/p.u. violation]. The savings are calculated relative to a baseline scenario with no flexibility in energy consumption or coordination. Private costs encompass energy bills and battery depreciation costs, without any additional reward payment for the cooperative voltage constraint management.

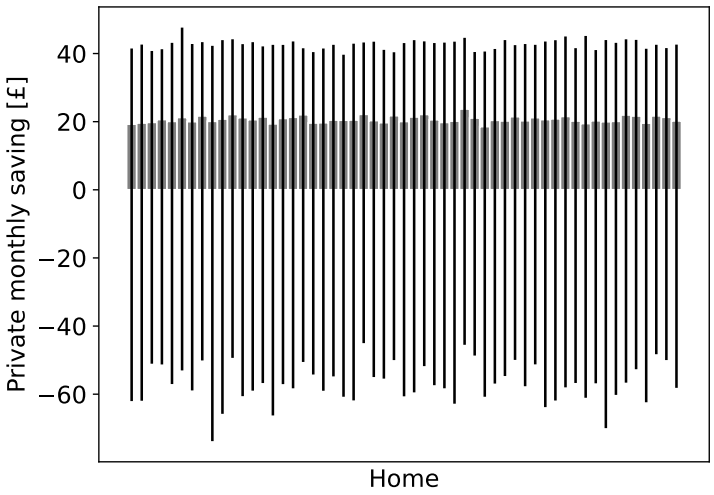


Figure 6.10: Private savings in energy bills and battery depreciation cost over one month for each of the 55 participants relative to the baseline inflexible scenario. Individual learned policies were tested in ten experiments. Cooperative setting with cooperative voltage management. The bar is the average value, and the black vertical lines are the spread from the minimum to the maximum monthly savings over the ten repeats.

Figure 6.10 illustrates that, although homes achieve savings on average even when cooperatively managing network constraints, the monthly private savings are more than twice as low (£19.81) as in the case without voltage management (£40.08) in Section 6.2.1. Moreover, all homes experience financial losses in the worst of the ten repeats relative to the case with no coordination, with the most significant individual monthly loss amounting to £73.78.

A comparison of the distributions of monthly savings for all homes, both with and without voltage management, is provided in Figure 6.11, demonstrating the substantial variability in savings when managing voltage constraints. While the standard deviation of monthly savings without voltage management is £12.58, equivalent to 31.4% of the mean saving, this standard deviation increases to £45.19, or 228.1% of the mean, in the case of voltage management.

The distribution of these savings (and losses) can be categorised into two groups. In eight out of ten repeated experiments[4], positive monthly savings are achieved,

[4]For each repeat, new policies were initialised, training and testing data were generated, and a new learning trajectory experiment was performed with different random seeds, thereby mitigating any potential bias introduced by the initial seed values.

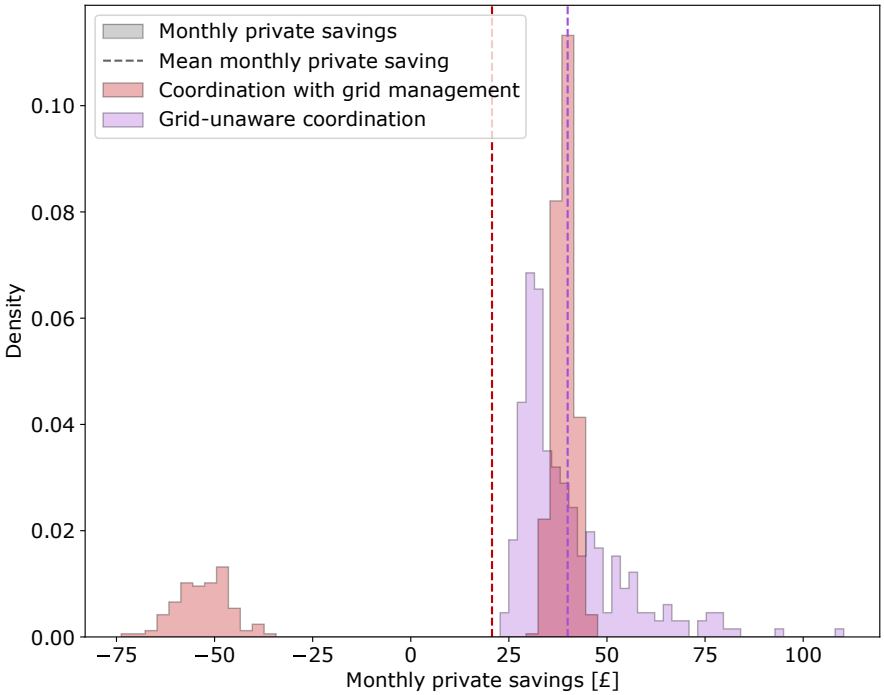


Figure 6.11: Distributions of private savings in energy bills and battery depreciation cost with and without cooperative voltage management.

with a mean of £39.07 and a standard deviation of £12.66, similar to the case without voltage management. No home ever experiences a monthly loss. In contrast, two repeats catastrophically fail to achieve satisfactory coordination, producing average savings of −£57.22 and a standard deviation of £46.41. 86.7% of the monthly savings over all homes and repeats are negative. It is noteworthy that, besides incurring financial losses for energy users relative to the baseline, the unsuccessful policies also fail to reduce the likelihood of voltage constraint violations. Thus, while successful repeats reduce the number of voltage constraint violations by 47.7%, the unsuccessful policies increase them by 122.2%.

A comparison between two successful and unsuccessful learning trajectories is presented in Figure 6.12.

Two key observations emerge when comparing policies without management of voltage constraints to policies successfully managing them. Firstly, policies that do not aim to manage constraints converge more rapidly than those that do. The actions agents select during training gradually converge towards a distribution of end

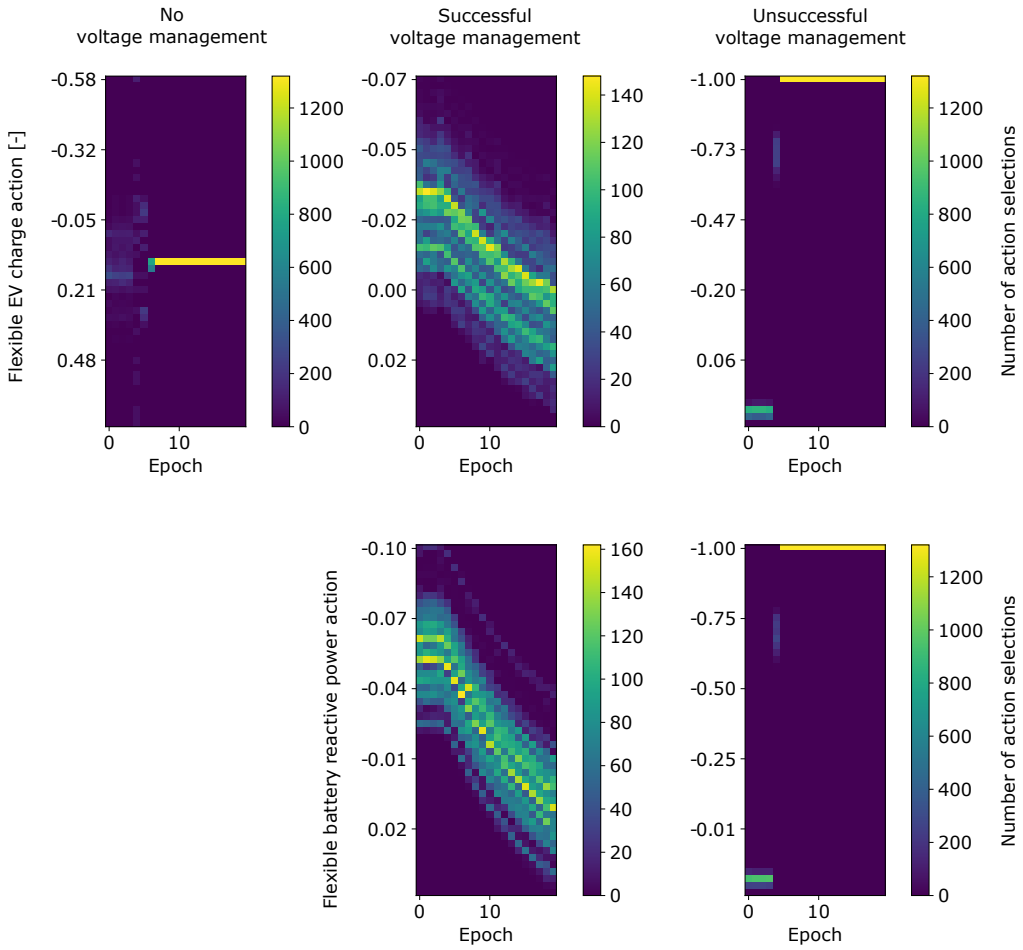


Figure 6.12: Learning trajectories in experiments with no voltage management, successful and unsuccessful voltage management. The heat map denotes the number of times each action value interval has been selected across all agents and repeats.

policies for each agent. Additionally, the resulting policies with voltage management produce actions with a standard deviation four orders of magnitude higher than those without. This increased variability in learned policies may help manage voltage constraints by enhancing the variability of behaviours in the grid. If all homes behaved deterministically and uniformly, the natural diversity on which the grid relies may be diminished, potentially overloading the network at certain times [182].

When in turn investigating the policies that fail to manage voltage constraints, there is no evidence of such a gradual learning as observed for successful policies. Instead, the failed policies diverge to extreme actions (-1 values plotted in Figure 6.12, other failed policies diverge to 1).

This failure to converge to a suitable policy in 20% of learning attempts, an outcome never observed in experiments without voltage management, indicates that the problem of learning cooperative voltage management is complex and can lead to instabilities in gradient updates. Neural networks become highly sensitive to initial conditions and data distributions, which can result in instability. If some value functions are drastically over- or under-estimated early in the learning process, promoting the exploration of a wider range of action values may be insufficient. Over time, these errors accumulate, and the policy may become trapped in a suboptimal configuration.

When implementing the unsuccessful policies over a month, the variability in savings can again be categorised into two groups. Some epochs and homes achieve positive savings, averaging £16.33, but 86.7% yield losses averaging £68.53. The distribution of savings and losses varies predominantly across epochs, with a standard deviation of £8.15 in average savings across epochs, whereas there is lower variation across homes, with a standard deviation of £1.81 across homes. In other words, all homes face similar risks of monetary losses at each epoch, as opposed to a few homes consistently sacrificing their financial utility for the coalition.

If deploying MARL-based cooperation frameworks with impacts on real-life electricity infrastructure, preliminary assessments of learned policies should be conducted to identify unstable policies before implementation. Additionally, in potential applications of the framework proposed in this thesis, compensatory payments may be necessary for users experiencing reduced savings as a result of cooperatively managing voltage constraints. An insurance mechanism could be implemented to distribute monthly savings more equitably, ensuring that no home is worse off in any given month. These considerations extend beyond the scope of this thesis.

Potential market designers may need to weigh the cost of such residential ancillary service provision relative to other ancillary service providers and to network reinforcement. While local residential energy users can provide highly localised services, such as mitigating resistive losses, addressing capacity limits of

network components, and managing voltage constraints, larger-scale assets may be more effective in providing more generic services, such as operating reserves or emissions reduction [22].

6.2.3 Other benefits

The savings for energy presented above do not include two further potential sources of monetary savings. Firstly, all energy users may experience reduced energy bills thanks to long-term reductions in network operation and investment costs resulting from the more effective coordination of DERs. Secondly, in the future, prosumers may become eligible for direct reward payments linked to their participation in post-operation incentive schemes. Examples of such schemes include potential markets for greenhouse gas emission reductions and voltage constraint management. While these markets are well-established for large industrial users, they are still in their early stages of development within the residential sector.

Moreover, a secondary objective in taking part in local markets is not only to minimise the expectancy of their energy costs but also to minimise their risk exposure through the coordination of resources to manage operational uncertainty [285]. The rewards obtained through cooperation are thus not only higher but also have higher reliability than the baseline, as illustrated in Figure 5.5.

Further shared co-benefits to the community include air quality improvement, and combating energy poverty [286].

Finally, beyond personal utility optimisation goals, distributed resource owners may also seek to contribute toward societal and community objectives, fulfil altruistic goals and fit within their community's social norms [287]. Communities may come together to pursue larger national policy goals, such as net-zero emission targets.

6.3 Robustness of positive impacts under variations of the implementation environment

Testing reinforcement learning policies in contexts that differ from the training data and environment is crucial for assessing their robustness and generalisability.

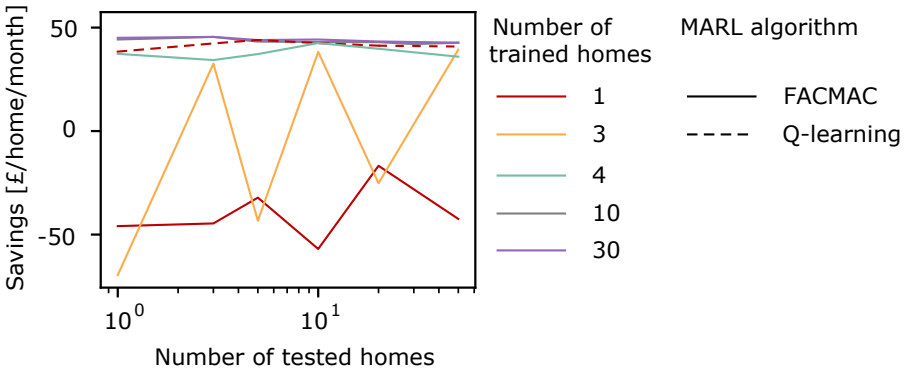


Figure 6.13: Savings as the number of tested homes increases for different numbers of trained homes and methodologies.

In real-world applications, AI agents must adapt to novel, unforeseen situations, and variations in environmental conditions. Evaluating policies in diverse scenarios helps enhance their practical utility and trustworthiness.

This section assesses the performance of trained policies in contexts varying from the training environment: in the presence of passive homes, distributing policies to a larger number of homes, and in different years.

6.3.1 Distributing pre-trained policies

Figure 6.13 investigates the potential for a set of policies to be distributed to a larger number of homes to reach coordination below a substation. The evaluation examines the savings per home as a fixed set of pre-trained policies is distributed to an increasing number of homes within a testing environment. In cases where there are more homes than pre-trained policies, the policies are randomly allocated to the additional homes in the testing environment. This analysis yields four primary insights.

Firstly, this figure reinforces the observation made in Chapter 5 that FACMAC is unable to learn effective policies for fewer than four homes. Even when the policies trained with fewer homes are distributed to a larger group, they cannot achieve cooperation. A minimum of four homes during the simulated training phase is necessary to combine a sufficient number of agent networks, leading to more reliable

estimates of the global critic. As recommended in Section 5.3.4, optimisation-informed IQL should be used for fewer than four homes coordinating their DERs.

Secondly, as was the aim of this analysis, this figure demonstrates that, given that the appropriate MARL methodology is accordingly selected, the trained policies can maintain their applicability and deliver value even in environments where the number of homes differs from the training. The policies are, by design, highly adaptable and resilient to structural changes in the environment, owing to their training in conditions characterised by high stochasticity and partial observability. This notion was exemplified in Section 4.3.4, which showcased how the MARL policies adopt robust and stable policies by smoothing out profiles while, in stark contrast to an omniscient optimiser that takes extremely precise actions based on an assumption of certainty.

Thirdly, the applicability of learned policies to be distributed to more homes helps elucidate one outstanding question around privacy: although the implementation of learned policies is achieved without sharing private data (see Chapters 3 to 5), does the training of individual policies require historical personal data? This analysis shows that policies trained with training data for a given set of homes can be implemented in numerous other homes and retain high coordination performance. Therefore, any given home neither shares its personal data in real-time during operational phases nor is obligated to provide its individual historical data for the purpose of training.

Lastly, this analysis paves the way for further advancements in the scalability of the MARL algorithms. This expansion encompasses not only the ability to scale up the training domain to accommodate 100 homes, as demonstrated in Chapter 5, but also the potential to distribute the trained policies to an even larger number of homes.

6.3.2 Interactions with uncoordinated homes

One fundamental question for the implementation of MARL policies in real electricity systems is whether the value of a simulated coalition of homes can materialise in the presence of other homes in the network. While the coalition operates in isolation

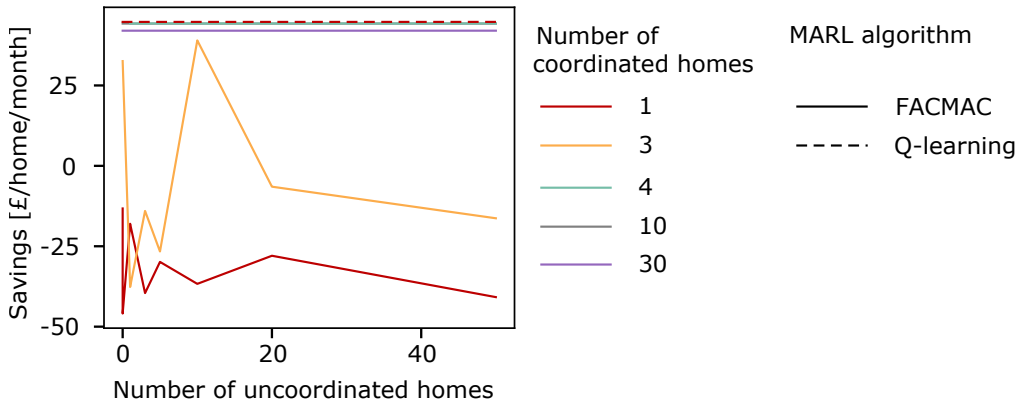


Figure 6.14: Savings per coordinated home as the number of passive homes in the environment increases for different numbers of trained homes and methodologies.

within the simulated environment, the objective function encompasses factors such as greenhouse gas emissions, upstream losses, and distribution network costs, which are influenced by non-participating homes.

Conversely, it is also important to ensure that the coordinated participants do not adversely affect the non-coordinating energy users. Not all energy users possess the capability or willingness to participate in local energy markets, and some may keep operating in a non-connected, non-flexible fashion. A significant risk lies in the possibility that local markets may increase the distribution network costs. The passive participants may disproportionately bear these costs if flexible resource owners shift away from traditional energy supply billing systems that usually settle network management costs [285]. Passive participants should, at the minimum, not be negatively affected by local markets and, ideally, gain co-benefits from these markets [288]. The coordination of energy users may for example reduce costs for all if it improves network constraint management.

Figure 6.14 compares the performance achieved by a fixed number of cooperating homes as the number of passive homes in the system increases.

The figure demonstrates that, provided the policies are trained using Q-learning for scenarios involving fewer than four homes, as recommended in Section 5.3.4, the trained policies are robust to variations in the number of non-coordinated energy users in the system. This finding reaffirms the insight derived in Section 6.3.1 that

the policies trained under conditions of high uncertainty display a high degree of adaptability to dynamic environmental changes.

Regarding the effects of coordination on uncoordinated homes, their incurred costs can be divided into two components: direct private costs and shared system costs. Firstly, the private costs of energy users, which encompass energy bills and battery depreciation, are independent of one another. These costs are solely influenced by local decisions regarding the timing of household flexible loads, heating, and the charging and discharging of EV batteries. Consequently, the private costs of uncoordinated energy users remain unaltered, regardless of the actions taken by the coalition.

Secondly, a case with 30 coordinated and 30 uncoordinated homes is investigated to assess the impact of coordination on shared system costs. Additional value is observed to be accrued beyond the private savings realised by the coordinated homes. Overall, an equivalent £4.81 per uncoordinated home is obtained in global value beyond their private savings. 59.2% of this shared value is attributed to reduced grid losses and 33.9% by reduced greenhouse gas emissions. While existing markets and mechanisms cannot directly capture this social value, it is indicative that the coalition generates an excess of social value rather than imposing negative externalities on the system.

6.3.3 Deploying policies in different years

Figure 6.15 illustrates the savings achieved by policies trained and tested in different years.

The weather, energy prices, and carbon intensity of grid electricity provision were varied accordingly. Crucially, Figure 3.12 illustrates how the historical electricity real-time prices used in the experiments varied between 2020 and 2022. Two discernible trends emerge from the data. Firstly, energy prices have increased yearly, with average prices of 9.34, 22.79 and 36.18 p/kWh, in 2020, 2021 and 2022, respectively. This constitutes year-on-year relative price increases of 144.2% and 58%. Secondly, concurrently with these price hikes, price curves deviate from the

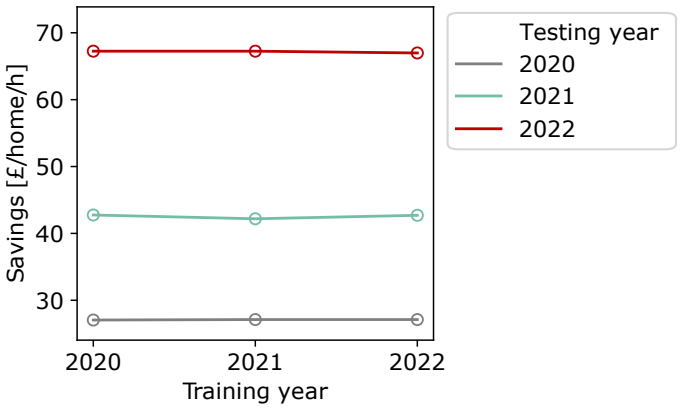


Figure 6.15: Savings as the year the policy was trained in changes for different testing years. Experiments using FACMAC. The shaded areas correspond to the interquartile range.

typical intra-day variability, as price values consistently reach the price cap. Despite these flatter profiles, which diminish the potential for energy arbitrage, Figure 6.15 distinctly shows that the predominant determinant of achievable savings is the grid prices in the testing or implementation environment.

Furthermore, this experiment underscores that policies learned at a given price level and profile pattern exhibit robust applicability in diverse contexts, further substantiating the results of the analyses in Sections 6.3.1 and 6.3.2. In the presented experiment, the normalised grid price state values that form the basis for selecting actions were simply normalised using the prevailing highest and lowest prices, and the policy directly applied.

6.4 Concluding discussion

This chapter has addressed the final sub-question investigated in this thesis, namely: *Can the algorithms yield positive impacts for energy users and the grid?*

This chapter reveals that, without explicit consideration of grid management, the MARL-based coordination of residential energy flexibility yields likelihood of voltage constraint violations 49.2% lower than the central, omniscient optimiser (Figure 6.2). Moreover, the cooperative management of voltage deviations in the electricity grid can mitigate the adverse effects associated with the growing

prevalence of DERs within the distribution network. As demonstrated in Figures 6.7 and 6.8, this MARL-based coordination effectively shaves both morning and evening peaks and yields 63.6% of the voltage constraint violation reduction achievable by a centralised, omniscient optimiser.

Furthermore, this chapter indicates that access to real-time network voltage information can enhance cooperative voltage constraint management outcomes relative to an identical scenario where this information was not available. As depicted in Figure 6.6, providing this information to agents alone can lead to a substantial reduction in the likelihood of voltage constraint violations by 53.0% relative to when this information is not available, given a voltage constraint violation penalty coefficient of $1{\times}10^2$ [£/p.u. violation]. Participants were thus found to be able to reduce the number of voltage constraint violations they cause by 43.1%% relative to an inflexible baseline, though not eliminating all under-voltages caused by the introduction of DERs into the distribution network.

Additionally, the analysis underscores the value that the MARL-based co-operation of residential energy flexibility can deliver to energy users. Without voltage cooperation, each prosumer readily saves 20.9% of their baseline private battery and energy costs. Moreover, the actions of the coalition do not impact the private costs of uncoordinated energy users, while concurrently generating social value through reductions in upstream grid energy losses and greenhouse gas emissions (Section 6.3.2).

Nonetheless, when managing network constraints, Figure 6.11 illustrates how private savings are reduced by half. Furthermore, the learning of policies becomes unstable and incurs monthly losses for energy users in 17.3% of cases. Consequently, while prosumers save money in expectancy by cooperating (Figure 6.9), future compensation mechanisms should be created to distribute the additional value generated equitably. In light of these instabilities, further research is required to assess whether fully decentralised MARL-based residential energy flexibility coordination is the most cost-effective and safest way to manage distribution network constraints and, if so, what market frameworks can incentivise cooperation.

Moreover, regulatory hurdles may have to be overcome, particularly in the context of the UK's less-defined regulatory framework compared to EU directives [289].

Finally, this chapter demonstrates that the policies learned in conditions of high stochasticity and partial observability are robust and adaptable to changes in environmental conditions. Figure 6.13 showcases the potential for a set of policies to be distributed to a larger number of homes to reach coordination below a substation, provided the adequate methodology is selected as per the recommendations of this thesis (Section 5.3.4). Given that optimisation-informed IQL is used for fewer than four coordinated homes, learned policies exhibit a degree of generalisability that accommodates varying numbers of testing homes, passive homes and deployment years.

Overall, this chapter underscores the potential of the MARL-based cooperation framework proposed in this thesis to deliver value to both the network and its users. Moreover, building upon the analysis of algorithm reliability presented in previous chapters (Sections 4.3.5 and 5.3.3), Section 6.3 reaffirms the robustness of MARL algorithms in the face of structural environmental changes. Evaluating the policies across diverse scenarios contributes to the development of more reliable and adaptable AI systems capable of effective and safe performance across a wide spectrum of circumstances. Such enhancements bolster their practical utility and instil trustworthiness, particularly when interfacing with human subjects and critical energy infrastructure.

Criterion addressed	**Verification**
Positive impact: The risks of implementing AI systems in the electricity grid presented in Section 1.2.4 should be evaluated. This cooperation could not increase the likelihood of voltage constraint violations in the distribution network nor increase private costs for individual energy users.	✓

Table 6.1: Final assessment of algorithm success criteria set in Section 1.3

7 Conclusions

Contents

This thesis considered the research question: can residential energy flexibility be coordinated without sharing private data at scale to achieve a positive impact on energy users in the electricity grid?

This is a crucial question, as climate change risks require urgent energy decarbonisation. This, in turn, means that energy uses should be electrified where possible and electricity provision decarbonised. In practice, relying on intermittent renewable energy sources for this purpose increases the need for demand-side flexibility. While the distributed energy resources (DERs) in the residential sector have a sizeable

potential to provide such demand-side flexibility, it is so far under-exploited due to costs, acceptability and technical hurdles.

This thesis therefore set out to develop coordination methodologies that minimise the communication and computational infrastructure requirements, avoid interfering with social practices, and preserve privacy. Moreover, it aimed to design coordination algorithms that are scalable both in terms of coordination performance (savings per home maintained at scale) and computationally (minimal and constant local computational burden for implementation, and only first-order growth in training requirements).

A literature review of the DER coordination and MARL fields identified that important aspects of the research questions remained. A systematic literature review of the DER coordination field showed that terminological ambiguity exists, which this thesis has sought to resolve by designing an exhaustive and mutually exclusive taxonomy. Coordination strategies were classified based on the level of agency of decision-making, the structure of information sharing, and the game type (competitive or cooperative). Upon mapping existing publications onto this framework, it became evident that implicit cooperation remains significantly under-researched despite exhibiting characteristics well-suited for addressing challenges in residential energy coordination. The second part of the literature review motivated the use of MARL for implicit cooperation. It presented a classification of three main MARL approaches to solving a decentralised partially observable Markov decision process (Dec-POMDP): independent learning, centralised multi-agent policy gradients, and value function factorisation. This categorisation informed the research direction for the rest of the thesis: both independent learning, which was the most straightforward and explainable approach, and value function factorisation, where individual value networks are factored into a global value, were investigated. Centralised multi-agent policy gradients were discarded as its state-action spaces grow exponentially with the number of agents, thus limiting scalability.

Three research gaps were identified in the literature review: benchmark environments for the MARL-based coordination of residential DER with a standard

residential energy data generation tool, the demonstration of the use of multi-agent reinforcement learning for fully decentralised implicit cooperation of residential DER, and an assessment of its potential impact on distribution networks and energy users.

The main research question was investigated by answering in turn four subsidiary research questions corresponding to these gaps. They are listed below, along with the key answers obtained.

7.1 Answers to the subsidiary research questions

7.1.1 Can one assess the efficacy of algorithms that coordinate residential energy flexibility?

Assessing coordination algorithms for residential energy flexibility is a critical aspect of this research. A modelling approach was required to address the challenges involved in trialling algorithms in the real world during the development phase, due to infrastructure criticality and the need to minimise interference with private activities.

Therefore, a comprehensive testing and benchmarking environment was developed in Chapter 3. This environment incorporates a model of a local energy system, encompassing intermittently available electric vehicles, heat pumps, and distribution network modelling (Section 3.1). Network modelling methodologies were selected and adapted to ensure scalability of the reinforcement learning training and testing, with relaxations of complex current constraints, and employing an iterative approach with linearised power flow equations.

In the context of this thesis, the testing environment sets a framework for the fully decentralised and lightweight implementation of coordination. The testing model operates with home-level control of existing DERs using only local data, without the sharing of personal information or behaviour. Agents select actions in response to existing energy tariffs in the form of one-way communication signals to the users. The home must only perform a feed-forward through the previously trained policy weights or look up the action in a Q-table, and convert the output to control actions given local information (Section 3.2.4). In this environment, the computational burden for the decentralised execution of RL policies therefore remains minimal

and constant by design, even as the system size increases. The control actions are obtained from local information in linear operations, which could be efficiently computed at the domestic level using inexpensive chips with minimal energy demands (less than 1 kilo floating point operations, kFLOPs). The execution part of the computational scalability requirements of this thesis (Section 1.3) is therefore met by design, while the training scalability will depend on algorithmic choices.

By employing cooperative training, the homes endeavour to maximise joint rewards rather than solely focusing on their individual rewards based on the received price signal. This testing environment provides the means to assess the efficacy of various coordination algorithms relative to lower and upper-bound benchmarks. The passive baseline serves as the lower bound, while a convex optimisation model acts as the upper bound.

Moreover, the timing required to train policies is recorded to assess computational scalability. Finally, the performance of RL algorithms is assessed not only in terms of average performance, but also in terms of variability and risk (Section 4.3.5).

Additionally, the testing model includes a generative adversarial network (GAN)-based data generator, known as HEDGE (Home Energy Data Generation), which generates semi-random data based on real-life datasets (Section 3.3). This data-driven approach enables a realistic representation of residential energy resources with adequate modelling of uncertainty for robust training and testing, while preserving magnitude and behavioural continuity for individual homes over time.

The HEDGE tool not only benefits this research but has the potential to be employed for various data-driven applications in the field of residential energy forecasting and control. The characterisation and simulation of residential energy resources are of increasing interest given their potential for demand-side response (Chapter 1). Particularly, data-driven methods are of particular interest in the field of residential energy forecasting and control, due to the high uncertainty at the local level, the limited availability of communication capability, and the limited scalability of centralised optimisation methods (Chapter 2).

By pre-training neural network weights for data generation, the HEDGE tool simplifies data preparation and training steps for researchers and practitioners, significantly reducing the required download size (from 40.12 GB to 125 kB), and cutting the data processing and neural network training computational needs. Only a feed-forward is required to obtain training and testing data.

Overall, the development and implementation of the testing and benchmarking environment, along with the HEDGE tool, successfully answer the first sub-research question. The testing framework provides a robust means of evaluating coordination algorithms against relevant benchmarks, and the HEDGE tool facilitates data-driven analysis and control within the residential energy sector, demonstrating its potential for various applications in the field. This meets the first requirement initially stipulated in Section 1.3.

7.1.2 Can one successfully coordinate residential energy flexibility without sharing private data?

Once the modelling and testing environment is established, residential energy coordination algorithms can be tested in computational experiments.

A new class of MARL-based implicit cooperation strategies is proposed for privacy preservation and scalability (Chapter 2). These strategies allow agents to learn RL policies using a data-based, model-free statistical approach, exploring a shared environment and interacting with decentralised partially observable Markov decision processes (Dec-POMDPs). The RL policies are trained to maximise global rewards in off-line simulations in the rehearsal phase, using historical data prior to online implementation. The pre-learned policies are then used to implicitly cooperate by making decisions under uncertainty given limited local information only.

Standard independent learners are found to be inadequate in achieving coordination under partial observability in a stochastic environment due to challenges like the Pareto selection problem, the stochasticity of the environment and the behaviour of other concurrently exploring agents, and the nonstationarity of the

environment (Section 4.3.3). A coordination mechanism is therefore proposed to allow the centralised rehearsal of decentralised cooperation.

The research proposed an *optimisation-informed* independent learning approach (Section 4.2.2). Agents are given access to the results of convex optimisation, allowing them to learn from an omniscient and time-travelling optimiser with perfect knowledge of the current and future system variables. This approach enables the agents to learn policies to be used under partial observability, aiming for actions that statistically perform well under uncertainty. The use of marginal rewards further enhanced learnability, allowing agents to assess the impact of their actions on global rewards.

The coordination mechanism achieved significant savings, obtaining 56.4% of the maximum total savings achieved by the omniscient and time-travelling optimiser on average (Section 4.3.3). The new class of optimisation-based learning outperforms environment-based learning across different numbers of homes, offering higher savings and a lower inter-quartile range. This coordination performance is maintained as the system size increases, measured in savings in energy, network and carbon costs obtained per home and month. Importantly, the proposed methodology successfully addressed acceptability issues by ensuring no interference in personal comfort, no communication of personal data, and control of appliances at the local level. Flexibility is utilised without hindering daily activities or requiring human participation. This meets the second requirement initially stipulated in Section 1.3.

At the end of Chapters 3 and 4, one question remains around privacy: although the implementation of learned policies is achieved without private data, does the training of individual policies require historical personal data? This is investigated in Chapter 6, which shows that policies trained with training data for a given set of homes can be implemented in numerous other homes and retain high coordination performance.

With this distinction elucidated, it can be posited that the examined methodologies effectively establish coordination while upholding privacy, not by relying

on robust encryption mechanisms, but by preventing the initial sharing of data outside of homes altogether.

In terms of computational scalability, the fixed-size Q-tables avoid the curse of dimensionality, as the state and action space size remains constant with the number of agents. This means that memory limitations were avoided. However, the generation of training data requires running optimisations on input data, and baseline simulations are required to compute marginal rewards. Although the computational time for pre-learning is not strictly a limiting factor as it is performed off-line ahead of implementation, the requirements grow at a second-order rate $O(n^2)$ with the number of homes (Section 4.3.6), which falls short of the third requirement for computational scalability requirements set out in this thesis. Therefore, an amended approach that meets both the performance and computational scalability requirements was next developed.

7.1.3 Can one achieve this in a computationally scalable manner?

Computational scalability is assessed in two phases: centralised training and decentralised execution. As decentralised execution scalability was already guaranteed by design, here we focus on the scalability of RL policy training.

In the centralised training phase, a deep multi-agent actor-critic reinforcement learning algorithm, known as Factored Multi-Agent Centralised Policy Gradients (FACMAC), is employed to improve scalability (Section 5.2.2). This algorithm combines the advantages of centralised multi-agent policy gradients and value function factorisation frameworks. By estimating a mixing network during training, agents can evaluate their impact on global rewards, reaping the benefits of a centralised critic to allow for more coordinated policy changes. This potential had thus far not been investigated for the coordination of residential DERs. We find that there is particularly high value in cooperation signals that take a global view of the system in this field, due to the global rewards being highly dependent on the cumulative impact of multiple agents taking actions simultaneously.

Deep reinforcement learning is used to handle continuous state and action spaces of higher dimensions and make decisions for the entire day ahead. Convolutional neural networks prove superior to linear ones by capturing patterns in data sequences. The use of a non-linear monotonic factorisation function estimator is also crucial in achieving coordination, as it uses a non-linear monotonic factorisation function estimator to estimate the global value, as opposed to simply adding up individual values (Section 5.3.2).

The approach demonstrates cooperation without the need for running optimisations for four homes and more (Section 5.3.2). Significant savings of £42.42 per home and month on average are achieved for energy users, the distribution network, and greenhouse gas emissions. The optimisation-informed independent learning and the centralised but factored critic approaches both provide a global coordination mechanism and achieve equivalent performance, while having structural differences that lead to varying computational efficiencies.

The reason for this large discrepancy in computation efficiency is that, in optimisations using the interior point method, the inverse of the Jacobian of the Karush–Kuhn–Tucker (KKT) conditions must be computed in each Newton-Raphson update step. As this includes derivatives with respect to all decision variables of the problem for all constraints and the objective function, the Jacobian grows with $O(n^2)$.

In FACMAC on the other hand, each back-propagation includes a Jacobian-gradient product of the value error with respect to the networks' weights for each operation in the graph. Analysis showed that the back-propagation from the centralised but factored critic mirrors global optimisations, while taking into account the partial observability of agents taking actions. It aims to update only the knowledge necessary to know which actions agents should take and when. This resulted in a coordination mechanism that achieved comparable performance to the optimisation-informed independent learning approach but with more efficient computational scaling. The number of value networks only grows linearly with

the number of homes. The corresponding linearly growing computational time requirement $O(n)$ was confirmed empirically.

Experiments confirmed this, and resulted in computational requirements 34 times lower for 30 agents (Section 5.3.2). This made the FACMAC approach computationally scalable for systems with a large number of homes (e.g. at the feeder level ∼100 homes), as per the third requirement of this thesis.

Although hyper-parameter tuning is more complex and explainability lower for FACMAC relative to IQL, this thesis therefore recommends using the optimisation-informed IQL algorithm for systems with fewer than four homes and the centralised but factored critic approach above this number to ensure computational scalability.

7.1.4 Can the algorithms achieve positive impacts for energy users and the grid?

Finally, Chapter 6 evaluates the impact of coordination on a simulated low-voltage network, and on energy users.

This chapter reveals that, without explicit consideration of grid management, the MARL-based coordination of residential energy flexibility yields likelihood of voltage constraint violations 49.2% lower than the central, omniscient optimiser (Figure 6.2). Moreover, the cooperative management of voltage constraints can mitigate the adverse effects associated with the growing prevalence of DERs within the distribution network. As demonstrated in Figures 6.7 and 6.8, this MARL-based coordination effectively shaves both morning and evening peaks and yields 63.6% of the voltage violation reduction achievable by a centralised, omniscient optimiser.

Furthermore, this chapter indicates that access to real-time network voltage information can enhance cooperative voltage constraint management outcomes relative to an identical scenario where this information was not available. As depicted in Figure 6.6, providing this information to agents alone can lead to a substantial reduction in the likelihood of voltage violations by 53.0% relative to when this information is not available, given a voltage violation penalty coefficient of $1{\times}10^2$ [£/p.u. violation]. Participants were thus found to be able to reduce

the number of voltage violations they cause by 43.1% relative to an inflexible baseline, though not eliminating all under-voltages caused by the introduction of DERs into the distribution network.

Additionally, the analysis provided underscores the value that the MARL-based cooperation of residential energy flexibility can deliver to energy users. Without voltage cooperation, each prosumer readily saves 20.9% of their baseline private battery and energy costs. Moreover, the actions of the coalition do not impact the private costs of uncoordinated energy users, while concurrently generating social value through reductions in upstream grid energy losses and greenhouse gas emissions (Section 6.3.2).

Nonetheless, Figure 6.11 illustrates how private savings are reduced by half when managing network constraints. Furthermore, the learning of policies becomes unstable and incurs monthly losses for energy users in 17.3% of cases. Consequently, while prosumers save money in expectancy by cooperating (Figure 6.9), future compensation mechanisms should be created to distribute the additional value generated equitably. In light of these instabilities, further research is required to assess whether fully decentralised MARL-based residential energy flexibility coordination is the most cost-effective and safest way to manage distribution network constraints and, if so, what market frameworks can be designed to incentivise cooperation.

Finally, this thesis demonstrated that the policies learned in conditions of high stochasticity and partial observability are robust and adaptable to changes in environmental conditions (Section 6.3). Figure 6.13 showcases the potential for a set of policies to be distributed to a larger number of homes to reach coordination below a substation, provided the adequate methodology is selected as per the recommendations of this thesis (Section 5.3.4). Given that optimisation-informed IQL is used for fewer than four coordinated homes, learned policies exhibit a degree of generalisability that accommodates varying numbers of testing homes, passive homes and deployment years.

Overall, the answer to this research question is that the MARL-based cooperation proposed in this thesis has the potential to deliver value to both the network and its users, with safeguards in place. Moreover, building upon the analysis of algorithm reliability presented in previous chapters (Sections 4.3.5 and 5.3.3), Section 6.3 reaffirms the robustness of MARL algorithms in the face of structural environmental changes. Evaluating the policies across diverse scenarios contributes to the development of more reliable and adaptable AI systems capable of effective and safe performance across a wide spectrum of circumstances. Such enhancements bolster their practical utility and instil trustworthiness, particularly when interfacing with human subjects and critical energy infrastructure.

7.1.5 Main conclusion

From the answers to the four sub-questions, this thesis clearly shows that residential energy flexibility can be coordinated without sharing private data at scale to achieve a positive impact for users and the grid, provided vulnerable users can be sheltered from risks of increased costs.

Specifically, this thesis has detailed two coordination mechanisms to overcome the challenges that independent learners face when seeking to coordinate residential DERs cooperatively in a decentralised manner. This thesis recommends using optimisation-informed independent Q-learning with marginal rewards for fewer than four homes, and a deep-learning actor-critic methodology with a centralised but factored centralised critic beyond. This can provide benefits for users, who save 20.9% of their baseline costs, and for the grid, as the cooperative management of voltage constraints can reduce the likelihood of violations by 43.1% relative to an uncoordinated baseline, though with trade-offs in reliability and energy user costs. The MARL policies are highly robust by design and this value is generalisable to varying environments.

7.1.6 Further insights

Moreover, beyond answering its core research question of this thesis, three insights that this thesis has released are:

1. One size does not fit all. There is no solution to rule them all, and it is crucial to adapt coordination methodologies to the specific context in which local energy coordination is being implemented. The coordination approach discussed in this dissertation is explicitly not an optimal one, but it serves as a niche application. Particularly, it can be applied when there is no capacity or preparedness for the systematic installation of ICT infrastructure, active human participation, and personal data sharing. In these contexts, the statistical approach developed in this thesis can improve outcomes where theoretically optimal, infrastructure-heavy approaches could not be implemented. However, in other contexts where these factors are feasible, alternative methodologies should be explored.

2. To connect this thesis within the broader context of DER coordination, it is important to recognise that, while this was an engineering thesis, the coordination of distributed energy resources is not just an engineering problem. The rise of local energy coordination will involve the collaboration and engagement of various stakeholders, including energy suppliers, network operators, regulators, participating and non-participating homes, coordinators, and communities. Preparing and aligning these stakeholders is needed to pave the way for implementing algorithms like those developed in this thesis. Previous pilot projects have encountered challenges and even had to be aborted due to the lack of preparedness from any of these stakeholders.

3. Finally, a motivation for this work was that, while long-term innovations like nuclear fusion should be supported and hold immense potential for transforming the world, waiting for these breakthroughs should not be an excuse for inaction. This research emphasises that we do not need to wait for

substantial investments, time-consuming infrastructure changes, or a complete societal shift in behaviours. We already possess resources that can be utilised to make a difference. This thesis has imagined using the existing walls and air in our homes, the cars serving our existing travel needs, and our existing energy tariffs to achieve a significant impact.

7.2 Limitations

The limitations of this research stem mainly from the fact that experiments were conducted using a simulated model, in lieu of in a real network and homes. As the statistician George Box coined: "All models are wrong, but some are useful".

The use of modelling was selected as (a) risks to people's lives and critical infrastructure have to be minimised during this methodological development exercise; (b) the scope of this thesis had to fit within the scope of an individual research project, whereas pilot projects in communities pose numerous organisational, regulatory, community, financial, commercial and engineering challenges requiring more extensive teams. Due to this lack of implementation and verification of the methodologies developed, uncertainty remains on their plausible use. Efforts in Chapter 6 to address some major questions on their implementation have nonetheless provided some reassurance.

The sources of the remaining uncertainty stem mainly from:

- Training and testing data: reinforcement learning is a data-driven method and, as such, results depend on data. Different policies and results may be obtained in different contexts.

- Modelling simplifications: computational scalability was often favoured over modelling accuracy for applicability in the data-driven methodologies developed in this thesis; all models should be validated with different types of homes and networks.

- Market assumptions: while this thesis has shown that residential DERs could achieve a positive impact on energy users, the network and the climate, there currently is no market for rewarding such cooperation. In current market frameworks, homes would seek to maximise their individual utility only, resulting in lower global utility. This thesis has only shown what *could* be.

7.3 Thesis contributions and associated publications

The four primary original and significant contributions of this dissertation are enumerated below.

1. **A testing environment for MARL-based coordination of residential energy**: The DER coordination benchmarking and testing environment for multi-agent reinforcement learning algorithms includes HEDGE, a home energy data generator which uses GANs to generate realistic household consumption, PV generation and travel data for use in machine learning application. Moreover, it includes inflexible and optimal benchmarks, as well as network modelling. The code for the environment[1] and the HEDGE tool[2] are available on GitHub. This work was submitted in a paper currently undergoing peer review:

    - **Charbonnier F,** Morstyn T, McCulloch MD. Home Electricity Data Generator (HEDGE): An open-access tool for the generation of electric vehicle, residential demand, and PV generation profiles [290]

2. **Optimisation-informed independent learning for scalable coordination under partial observability**: A novel class of decentralised flexibility coordination strategies, optimisation-informed independent learning, allows for implicit cooperation with no communication of personal data. Agents under partial observability learn from omniscient, convex optimisations prior

[1] `https://github.com/floracharbo/MARL_local_electricity`
[2] `https://github.com/floracharbo/hedge`

to implementation for convergence to robust cooperation at scale. Moreover, fixed-size Q-tables mitigate the curse of dimensionality. This was published as:

- **Charbonnier F,** Morstyn T, McCulloch MD. Scalable multi-agent reinforcement learning for distributed control of residential energy flexibility. Applied Energy 2022;314:118825.[213]

3. **A demonstration of the use of a factored but centralised critic for computationally scalable DER coordination**: Deep multi-agent reinforcement learning with a centralised but factored critic can further improve the computational scalability of decentralised DER coordination for larger numbers of homes. This was presented in a paper currently undergoing peer review:

   - **Charbonnier F**, Peng B, Morstyn T, Vienne J, Stai E, Stanojev O, Hug G, McCulloch MD. Centralised rehearsal of decentralised cooperation: Multi-agent reinforcement learning for the scalable coordination of residential energy flexibility.

4. **The analysis of network impacts of MARL-based implicit coordination**: The impacts of the integration at scale of DERs with and without MARL-based implicit cooperation were analysed. These included the impacts on a simulated standard low-voltage network and energy users. This was also presented in the same paper currently undergoing peer review:

   - **Charbonnier F**, Peng B, Morstyn T, Vienne J, Stai E, Stanojev O, Hug G, McCulloch MD. Centralised rehearsal of decentralised cooperation: Multi-agent reinforcement learning for the scalable coordination of residential energy flexibility.

Moreover, two secondary contributions were:

1. **The provision of a unified taxonomy for the field of DER coordination**, which synthesises distributed energy resources coordination strategies to help resolve the terminological ambiguity in the field. This was published as: **Charbonnier F**, Morstyn T, McCulloch MD. Coordination of resources at the edge of the electricity grid: Systematic review and taxonomy. Applied Energy 2022;318:119188. [243]

2. **A book chapter** was published as: **Charbonnier, F**, Morstyn, T., McCulloch, MD. (2023). Active Players in Local Energy Markets. In: Shafie-khah, M., Gazafroudi, A.S. (eds) Trading in Local Energy Markets and Energy Communities. Lecture Notes in Energy, vol 93. Springer, Cham. [285]

7.4 Future research

Both the methodological research and the provision of the testing framework presented in this dissertation can serve as a foundation for further research at the intersection of MARL and DER coordination. Potential future research stemming from this work could thus seek to answer the three following research questions:

1. Can local forecasts of loads, PV generation, heating requirements and EV travelling patterns be generated as inputs for MARL-based control? How can control algorithms be adapted to improve robustness to prediction errors?

   Future research could investigate the feasibility and accuracy of local forecasts encompassing a range of critical parameters such as electrical loads, PV generation patterns, heating requirements, and EV travelling patterns. The primary aim would be to use machine learning techniques leveraging historical data. Moreover, research could focus on enhancing the robustness of MARL-based control systems in the face of prediction errors, which are inherent in any forecasting process. This may involve the development of reinforcement learning techniques that are inherently more resilient to prediction inaccuracies, or the incorporation of adaptive control mechanisms that can dynamically adjust to changing forecast conditions. This line of inquiry could help bridge

the gap between prediction and control, ultimately driving innovations that benefit both the academic community and real-world energy practitioners.

2. Can other MARL algorithms be implemented in the residential DER coordination testing environment and provide superior coordination and scalability performance?

   One could evaluate whether other MARL algorithms [223], such as Independent synchronous Advantage Actor-Critic (IA2C), Independent Proximal Policy Optimisation (IPPO), Multi-Agent Deep Deterministic Policy Gradients (MADDPG), Multi-Agent Advantage Actor-Critic (MAA2C), and Multi-Agent Proximal Policy Optimization (MAPPO) can be successfully implemented within the residential DER coordination testing environment presented in this thesis. Key performance indicators will include not only coordination performance but also computational scalability, data requirements, convergence, training stability and generalisation among others.

3. Can graph neural networks (GNNs) matching distribution network grid structure be used as mixing networks to enhance the learning of cooperative grid management?

   Research could be conducted to investigate the feasibility and effectiveness of adapting GNNs to represent the topological and operational characteristics of distribution networks. By aligning the GNN structure with the underlying grid structure, the network is anticipated to inherently incorporate the spatial and temporal dependencies critical for effective cooperative grid management. Topology-informed GNNs have thus previously been used to predict the optimal solutions of real-time ac-OPF problem [291]. Research could focus on how to adapt and extend this to a MARL framework.

4. Can remuneration frameworks be designed that are statistically game theory-compliant to incentivise MARL-based cooperation?

Existing market mechanisms must be accompanied by mechanisms to incentivise prosumers' cooperation to contribute towards a social value. Previous research has identified desirable game-theoretic properties of potential DER coordination remuneration mechanisms, including efficiency, incentive compatibility, budget balance, and group rationality [25, 76]. Further research is needed to extend this theory to a statistical framework to bridge the gap between cooperative MARL and game theory. Novel remuneration frameworks could be introduced that align individual agents' self-interests with the collective goals of the system, leading to equilibrium states where agents find it in their best interest to cooperate, even when faced with complex, dynamic, and uncertain environments.

5. Can worst-case violations be minimised and worst-case performance guarantees be achieved? Conventional ML training processes prioritise maximising the average performance over reliability, without providing guarantees on worst-case estimation errors. This limitation poses a substantial barrier to the adoption of ML in safety-critical systems, notably within power systems. One potential avenue for future work would be to leverage the insights and methodologies presented in [292–295] to be applied to the MARL framework developed in this thesis. These aim to concurrently optimise average performance and minimise worst-case violations, and with guarantees on the worst-case scenarios. This is a critical consideration for fostering trust and ensuring reliability in algorithms designed to manage distribution network constraints.

Appendices

A

Literature review additional material

A.1 Scopus search query

We systematically review the literature using a structured topic search query in the Scopus search engine [103], the largest abstract and citation database of peer-reviewed literature. We aim to select literature that lies at the intersection of the concepts of coordination, grid-edge participation and electric resources. The following sequence is followed to define the query terms for the literature search:

1. Select terms associated with each of the concepts of coordination, grid-edge participation and electric resources. Where relevant, only the root of each term is selected so that associated adjectives, nouns and verbal forms may also be included in the search.

2. Conduct a search to obtain all publications containing at least one of the terms associated with each category in their title or abstract.

3. Inspect the titles and abstracts of the first 50 results, adding missing terms encountered that are relevant to any of the three categories, and excluding irrelevant search terms, topic areas and journals from future searches.

4. Repeat (2) and (3) until the first 50 results are either relevant or do not contain terms that may be excluded from the search without excluding more relevant results.

The resulting query terms are presented in Figure A.1. Terms which were excluded from the corpus are listed in Table A.1. The body of literature obtained comprised of publications published between 1962 and 2020. Inspecting the number of publications over time in Figure 2.1, we further focus our literature review on the period between 1995, the year in which the body of literature started growing consistently, and 2020, resulting in 73,053 titles and abstracts. As expected, given current electric grid transformations, it is immediately apparent that the coordination of distributed grid-edge electric resources is an increasingly important topic of research. Publications including selected search themes rose exponentially from 248 in 1995 to 9,156 in 2020, doubling approximately every 4.7 years.

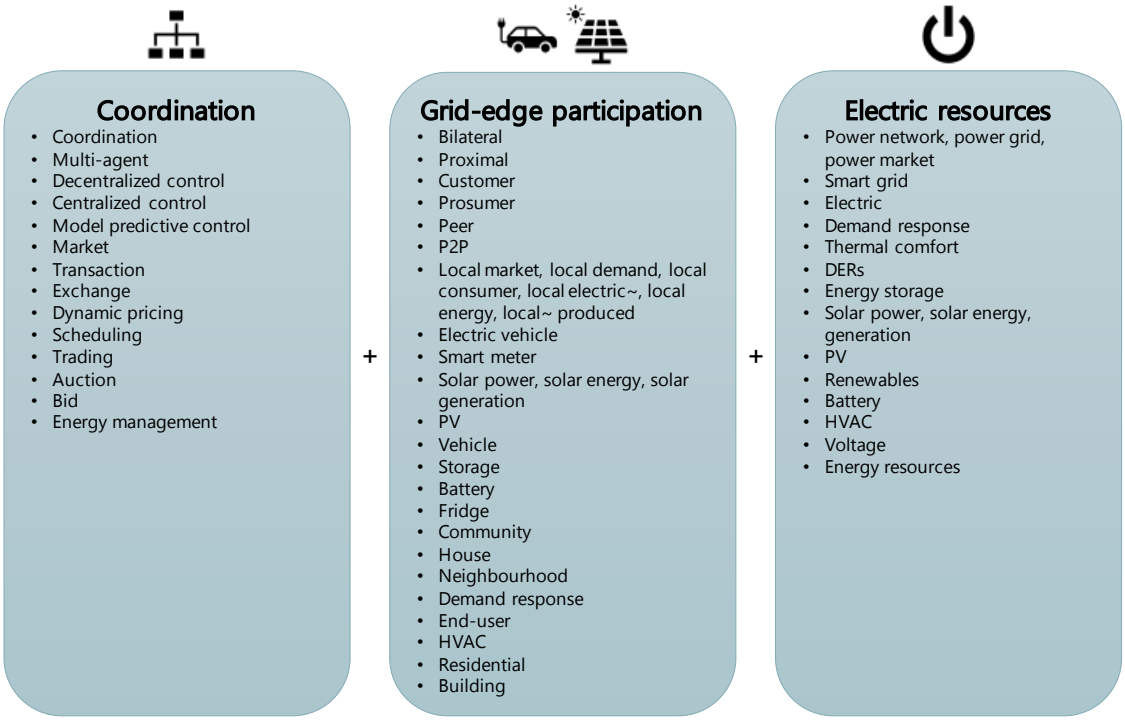


Figure A.1: Terms for the structured search query, associated with the concepts of coordination, grid-edge participation and energy resources.

Type	Key words
Title or abstract	cloud, gas, forestry, data center, optical, water supply, mechanics, life cycle assessment, Micro Phasor Measurement, tyre, manufacturing, clutchless, concrete, methane, satellite, torque, health, speech
Journal	chemistry, social science, materials, environment, sustainable production and consumption, Vibration and Acoustics, nano energy, Future Generation Computer Systems, chemical engineering, food, surface science, Minerals Engineering
Subject Area	Medicine, Social sciences Arts and humanities, Environmental science, Mathematics, Agricultural and biological sciences, Physics and astronomy, Earth and planetary sciences, Biochemistry, genetics and molecular biology, Materials science, Psychology, Chemistry, Chemical engineering, Immunology and biology, Pharmacology, toxicology and pharmaceutics, Neuroscience, Nursing, Health professions, Veterinary, Dentistry

Table A.1: Words excluded from topic search.

A.2 Systematic literature review: themes identification

Next, we identify research themes within this body of literature. Key words are identified from the first 50 titles and abstracts, and grouped in themes as listed in Table A.2. The absolute and relative prevalence of these major themes are displayed on Figure 2.3, so that trends may be identified. While the prevalence of themes in absolute numbers is highly dependent on the boundary of the body of literature defined in Appendix A.1, we identify research trends in relative terms by looking at prevalence over time, i.e. the share of publications within the body of literature that refers to each theme a given year. Moderate, fast and very fast increases correspond to average prevalences in the first half of the period (1995-2007) of less than 100%, two thirds and one third of that in 2020 respectively. In the rest of this section, we will individually describe the 17 major themes identified listed in Figure 2.3.

	Theme	Key words
1	Storage	storage, battery

2	Renewables	renewables, photovoltaic, PV, wind, hydro, solar power, solar generation, solar energy
3	Optimisation	optimise, maximise, minimise
4	Market	market, pricing, transaction, incentives
5	Network constraints	constraints, voltage, frequency, line capacity, stability, ancillary services, utility services
6	Uncertainty	uncertainty, risk, robust, stochasticity, forecasting, probabilistic
7	Demand Response	demand response, DR, demand side, customer side, flexible loads, flexibility, deferrable, load shifting
8	Residential sector	resident, domestic, house, household, home, community, neighbourhood
9	Electric vehicles	electric vehicle, EV
10	Decentralisation	decentralised, distributed
11	Thermal loads	thermal, temperature, heating, boiler, cooling, fridge, air conditioning, HVAC
12	Machine Learning	machine learning, ML, reinforcement learning, RL, artificial intelligence, AI, data-driven, neural network
13	Forecasting	forecast
14	Peer-to-peer	peer-to-peer, P2P, bilateral, matching
15	Game theory	game theory, game-theoretic, cooperation, coalition
16	Privacy	privacy, data protection, personal data, attack
17	Blockchain	blockchain, distributed ledger, DLT

Table A.2: Research themes emerging from the keyword search.

A.3 Extended detailed literature review

Paradigm	Ref.	Strategy	Unit type(s): Local generation	Unit type(s): Storage	Unit type(s): Thermal control	Unit type(s): Generic flexible load	Methods: RL use	Methods: Network constraints	Methods: Networks losses	Objectives: Operation costs	Objectives: Investment costs	Objectives: Battery degradation	Objectives: Renewables curtailment	Objectives: Power losses	Objectives: Energy arbitrage, peak shaving	Objectives: Load shifting, ancillary services	Objectives: Prosumer utility
Direct control	84	"event-based DR"															
	89	"centralised control strategy"		↘													
	102	"optimal power flow"	↘	↘				↘									
	102	"direct load control"	↘	↘				↘									
	100	"centralised control"															
	112	"multi-agent system"	↘	↘		↘											
	115	"top-down switching"															
	115	"centralized optimization"															
	126	"operator instructions"															
	67	"centralised dispatch"	↘	↘		↘		↘	↘								
	143	"model predictive control"	↘	↘				↘	↘					↘			
	144	"stochastic optimization framework"						↘							↘	↘	
	145	"model predictive control"	↘	↘	↘	↘		↘	↘	↘			↘		↘		
	146	"centralized model predictive control scheme"		↘				↘	↘			↘	↘				

Paradigm	Ref.	Strategy	Unit type(s)				Methods			Objectives								
			Local generation	Storage	Thermal control	Generic flexible load	RL use	Network constraints	Networks losses	Operation costs	Investment costs	Battery degradation	Renewables curtailment	Power losses	Energy arbitrage, peak shaving	Load shifting, ancillary services	Prosumer utility
	147	"optimal energy balance methodology"	↘	↘	↘					↘	↘						
	149	"consumer automated energy management system"	↘			↘	↘			↘							↘
	150	"fridge controller"			↘											↘	
	151	"direct load control architecture"			↘											↘	
Mediated competition	84	"organized markets"															
	84	"demand reduction bids"															
	30	"adaptive consumption-level pricing scheme"								↘							↘
	96	"coordinated market"															
	96	"community market"															
	98	"explicit demand response"			↘											↘	
	102	"demand response"	↘	↘				↘									
	102	"auction-based methods"	↘	↘				↘									
	114	"transactive energy service system"											↘				

Paradigm	Ref.	Strategy	Objectives: Prosumer utility	Objectives: Load shifting, ancillary services	Objectives: Energy arbitrage, peak shaving	Objectives: Power losses	Objectives: Renewables curtailment	Objectives: Battery degradation	Objectives: Investment costs	Objectives: Operation costs	Methods: Networks losses	Methods: Network constraints	Methods: RL use	Unit type(s): Generic flexible load	Unit type(s): Thermal control	Unit type(s): Storage	Unit type(s): Local generation
	115	"transactive control"	↘														
	116	"market-based multi-agent system"			↘								↘	↘		↘	
	117	"market-based control"	↘							↘			↘			↘	
	119	"market-based EV charging coordination"	↘	↘									↘			↘	
	120	"reinforcement learning-based dynamic pricing algorithm"	↘										↘	↘			
	121	"system-centric"		↘						↘	↘	↘					
	122	"continuous double auction"	↘	↘	↘	↘				↘	↘	↘				↘	↘
	123	"P2P local electricity market model"	↘	↘						↘		↘		↘	↘		
	124	"energy blockchain in microgrids"	↘	↘						↘	↘	↘				↘	↘
	125	"decentralized markets for distribution system flexibility"	↘	↘						↘	↘	↘		↘		↘	↘
	126	"price-responsive modes"	↘	↘													
	132	"P2P trading"	↘	↘		↘	↘			↘							↘
	67	"unidirectional pricing"	↘	↘		↘		↘		↘	↘	↘		↘		↘	↘

Paradigm	Ref.	Strategy	Unit type(s)				Methods			Objectives								
			Local generation	Storage	Thermal control	Generic flexible load	RL use	Network constraints	Networks losses	Operation costs	Investment costs	Battery degradation	Renewables curtailment	Power losses	Energy arbitrage, peak shaving	Load shifting, ancillary services	Prosumer utility
	140	"automatic P2P energy trading model based on reinforcement learning"	↘	↘									↘				
	152	"incentive-based demand response"				↘	↘									↘	↘
	153	"agile demand response"				↘	↘			↘							↘
	154	"Stackelberg approach"				↘											↘
	155	"distributed demand response"				↘	↘			↘					↘		↘
	156	"bilateral trading electricity market"	↘			↘										↘	↘
	157	"deep reinforcement learning for strategic bidding"	↘			↘	↘										↘
	158	"learning based bidding strategy"			↘		↘										↘
	35	"indirect customer-to-customer energy trading"					↘			↘					↘	↘	↘
Mediated cooperation	37	"energy collectives"	↘			↘				↘							↘
	76	"prosumer coalitions with energy management"		↘		↘											↘

Paradigm	Ref.	Strategy	Unit type(s): Generic flexible load	Unit type(s): Thermal control	Unit type(s): Storage	Unit type(s): Local generation	Methods: RL use	Methods: Network constraints	Methods: Networks losses	Objectives: Operation costs	Objectives: Investment costs	Objectives: Battery degradation	Objectives: Renewables curtailment	Objectives: Power losses	Objectives: Energy arbitrage, peak shaving	Objectives: Load shifting, ancillary services	Objectives: Prosumer utility
	33	"community-based market"	↘			↘				↘							↘
	118	"deep transfer Q-learning"			↘		↘	↘							↘		
	127	"multi-agent residential demand response"			↘		↘						↘		↘		
	128	"decentralized learning-based multi-agent residential DR"			↘		↘	↘								↘	↘
	129	"extended joint action learning"	↘	↘			↘	↘								↘	↘
	130	"P2P trading"													↘	↘	↘
	138	"distributed price-directed optimization"			↘	↘				↘		↘		↘			↘
	160	"semiautonomous operation"		↘											↘	↘	
	161	"coordinated multilateral trades"						↘	↘						↘	↘	↘
	162	"distribution locational marginal costs and hierarchical decomposition"			↘	↘				↘						↘	
Bilateral	96	"decentralized market"															

Paradigm	Ref.	Strategy	Unit type(s)				Methods			Objectives						
			Local generation	Storage	Thermal control	Generic flexible load	RL use	Network constraints	Networks losses	Operation costs	Investment costs	Battery degradation	Renewables curtailment	Power losses	Energy arbitrage, peak shaving	Load shifting, ancillary services	Prosumer utility
competition	102	"matching theory-based methods" for "P2P"	↘	↘				↘									
	121	"peer-centric"													↘		
	126	"bilateral transactive bids"															
	131	"prosumer centric local energy market"														↘	
	132	"P2P trading"	↘														
	133	"blockchain-based distributed double auction trade"	↘					↘	↘								↘
	134	"bilateral contract networks"	↘	↘		↘											↘
	67	"P2P energy trading"	↘	↘		↘											
	171	"bilateral contracts"															↘
	234	"P2P trading"	↘	↘		↘			↘						↘		
Bilateral	89	"distributed multi-agent control strategy"		↘													
	100	"distributed control"															

Paradigm	Ref.	Strategy	Unit type(s)				Methods			Objectives							
			Local generation	Storage	Thermal control	Generic flexible load	RL use	Network constraints	Networks losses	Operation costs	Investment costs	Battery degradation	Renewables curtailment	Power losses	Energy arbitrage, peak shaving	Load shifting, ancillary services	Prosumer utility
cooperation	102	"distributed methods"	↘	↘				↘									
	33	"full P2P market"								↘							
	113	"decentralized bilateral energy trading"						↘									↘
	67	"distributed dispatch"	↘	↘		↘											
	135	"parallel transfer learning"		↘			↘								↘	↘	
Implicit competition	84	"rate-based or price DR programs"															↘
	98	"implicit demand response"			↘										↘		
	102	"uncoordinated approaches"	↘	↘				↘									
	115	"price-reactive systems"															↘
	126	"autonomous mode"															↘
	176	"load scheduling algorithm"				↘										↘	↘
	177	"deep reinforcement learning based energy storage arbitrage"		↘			↘					↘					↘

Paradigm	Ref.	Strategy	Objectives: Prosumer utility	Objectives: Load shifting, ancillary services	Objectives: Energy arbitrage, peak shaving	Objectives: Power losses	Objectives: Renewables curtailment	Objectives: Battery degradation	Objectives: Investment costs	Objectives: Operation costs	Methods: Networks losses	Methods: Network constraints	Methods: RL use	Unit type(s): Generic flexible load	Unit type(s): Thermal control	Unit type(s): Storage	Unit type(s): Local generation
	178	"energy storage arbitrage in real-time markets via reinforcement learning"	↘										↘			↘	
	179	"device-based reinforcement learning"											↘	↘			
	180	"reinforcement learning controller"											↘		↘		
	181	"reinforcement learning control"	↘	↘			↘						↘		↘		↘
Implicit cooperation	89	"decentralised control strategy"		↘												↘	
	100	"autonomous control"		↘													
	136	"prediction-based multi-agent reinforcement learning"			↘								↘			↘	
	160	"fully autonomous operation"		↘											↘		
	137	"multi-agent reinforcement learning"									↘	↘	↘		↘	↘	↘

Paradigm	Ref.	Strategy	Unit type(s)				Methods			Objectives					
			Generic flexible load	Thermal control	Storage	Local generation	Networks losses	Network constraints	RL use	Prosumer utility	Load shifting, ancillary services	Energy arbitrage, peak shaving	Power losses	Renewables curtailment	Battery degradation	Investment costs	Operation costs

Table A.3: Distributed energy resources coordination strategies mapped onto the proposed taxonomy. Ticks indicate which controlled resources are specifically modelled (local generation, storage, thermal control, generic flexible load), whether reinforcement learning (RL) is being used as an example of coordination tool, and whether the impact of coordination strategies on network constraints and losses is modelled. In this table, energy arbitrage may refer to displacing energy use to times of lower prices or to times with higher renewable generation. Ancillary services may refer to the provision of reactive power and real-time management of network constraints, including voltage and frequency control. No ticks are indicated where an individual article does not specify the coordinated unit type or specific objective.

B

Heating model derivation

We use the simple hourly method heating model laid out in [237].

The input data used in the heating model in this paper is tabulated below. Note that these heating model and input data are meant as a generic building example that can be used to test the relative performances of the MARL coordination algorithms. Models and parameters used for the detailed study of a specific building should be validated with experimental data.

Symbol	Definition	Value	Unit	Reference
A_{d}	door area	1.4 × 2	m^2	
A_{f}	floor area	76	m^2	
A_{wd}	window area	1.4×1.4	m^2	
e	shielding coefficient	0.03	-	[296]
h	height of rooms	2.4	m	
h_{is}	heat transfer coefficient between the air node θ_{air} and the surface node θ_{s}	3.45	$\mathrm{W.m^{-2}.K^{-1}}$	[237]
h_{ms}	heat transfer coefficient between nodes m and s	9.1	$\mathrm{W.m^{-2}.K^{-1}}$	[237]
k_{party}	fraction of floor space that is party floor rather than on ground, for a one-storey building	0.5	[-]	
n_{min}	minimum external air exchange rate per hour for a habitable room	0.5	$\mathrm{h^{-1}}$	[296]
n_{50}	Air exchange rate resulting from a pressure difference of 50 Pa between the inside and the outside of the building, including the effects of air inlets, medium construction family dwelling	6	$\mathrm{h^{-1}}$	[296]
U_{g}	U-value for ground	1.0	$\mathrm{W.m^{-2}.K^{-1}}$	[297]
U_{r}	U-value for roof	1.0	$\mathrm{W.m^{-2}.K^{-1}}$	[297]
U_{w}	U-value for walls, ceiling against outside	1.5	$\mathrm{W.m^{-2}.K^{-1}}$	[297]
U_{wd}	U-value for windows	4.3	$\mathrm{W.m^{-2}.K^{-1}}$	[297]
ϵ	height correction factor	1	-	[296]
Λ_{at}	dimensionless ratio between the internal surfaces area and the floor area	4.5	[-]	[237]
τ	time step	3600	s	
COP	Coefficient of performance	3	[-]	

Table B.1: Input parameters to the heating model

We obtain the following intermediate parameter values:

- Effective mass area A_{m} [m^2] for a medium-class building [237]:

$$A_{\mathrm{m}} = 2.5A_{\mathrm{f}} \tag{B.1}$$

- Internal heat capacity of the building zone for medium-class building J.K^{-1} [237]:

$$C_{\mathrm{m}} = 165,000 A_{\mathrm{f}} \tag{B.2}$$

- Area of all surfaces facing the building zone A_{tot} [m^2] [237]:

$$A_{\mathrm{tot}} = \Lambda_{\mathrm{at}} A_{\mathrm{f}} \tag{B.3}$$

- The coupling conductance [W.K^{-1}] [237]:

$$H_{\mathrm{tr,is}} = h_{\mathrm{is}} A_{\mathrm{tot}} \tag{B.4}$$

- The coupling conductance between nodes m and s [W.K^{-1}] [237]:

$$H_{\mathrm{tr,ms}} = h_{\mathrm{ms}} A_{\mathrm{m}} \tag{B.5}$$

- Wall area (excluding windows and doors)

$$A_{\mathrm{w}} = 4\sqrt{A_f} h - 8A_{\mathrm{wd}} - A_{\mathrm{d}} \tag{B.6}$$

- The thermal transmission coefficient of walls [W.K^{-1}] [237]:

$$H_{\mathrm{tr,w}} = A_{\mathrm{w}} U_{\mathrm{w}} \tag{B.7}$$

- The thermal transmission coefficienf of the roof [W.K^{-1}] [237]:

$$H_{\mathrm{tr,r}} = A_{\mathrm{f}} U_{\mathrm{r}} \tag{B.8}$$

- The thermal transmission coefficient of the floor [W.K^{-1}] [237]:

$$H_{\mathrm{tr,f}} = A_{\mathrm{f}} (1 - k_{\mathrm{party}}) U_{\mathrm{g}} \tag{B.9}$$

- The heat transfer coefficient for opaque elements $H_{\mathrm{tr,op}}$ [W.K^{-1}] [237]:

$$H_{\mathrm{tr,op}} = H_{\mathrm{tr,w}} + H_{\mathrm{tr,r}} + H_{\mathrm{tr,f}} \tag{B.10}$$

- The opaque heat transfer coefficient is split between conductance transfer and $H_{\text{tr,em}}$ [237]:

$$H_{\text{tr,em}} = \frac{1}{\frac{1}{H_{\text{tr,op}}} - \frac{1}{H_{\text{tr,ms}}}} \tag{B.11}$$

- The thermal transmission coefficienf of windows [W.K^{-1}] [237]:

$$H_{\text{tr,wd}} = (A_{\text{wd}} + A_{\text{d}})U_{\text{wd,eff}} \tag{B.12}$$

- The conditioned air volume [m^3]

$$V_{\text{r}} = A_{\text{f}}h \tag{B.13}$$

- The hygiene minimum air flow rate of a heated space V_{min} [m^3.h^{-1}] [296]:

$$V_{\text{min}} = n_{\text{min}}V_{\text{r}} \tag{B.14}$$

- The infiltration through building envelope V_{inf} [m^3.h^{-1}] [296]:

$$V_{\text{inf}} = 2V_{\text{r}}n_{50}e\epsilon \tag{B.15}$$

- The air flow rate of heated space [m^3.h^{-1}] [296]:

$$V = \max(V_{\text{min}}, V_{\text{inf}}) \tag{B.16}$$

- The heat transfer by ventilation H_{ve} [W.K^{-1}] [296]:

$$H_{\text{ve}} = 0,34V \tag{B.17}$$

- The effective window U-value, corrected for the assumed use of curtains [W.m^{-2}.K^{-1}] [237]:

$$U_{\text{wd,eff}} = \frac{1}{\frac{1}{U_{\text{wd}}} + 0.04} \tag{B.18}$$

- Three helper transmission coefficients [237]:

$$H_{\text{tr,1}} = \frac{1}{\frac{1}{H_{\text{ve}}} + \frac{1}{H_{\text{tr,is}}}} \tag{B.19}$$

$$H_{\text{tr,2}} = H_{\text{tr,1}} + H_{\text{tr,w}} \tag{B.20}$$

$$H_{\text{tr,3}} = \frac{1}{\frac{1}{H_{\text{tr,2}}} + \frac{1}{H_{\text{tr,ms}}}} \tag{B.21}$$

- The heat flow rate from internal heat sources Φ_{int} [W] is taken as the sum of the average heat flow rate from appliances $\Phi_{\text{int,A}}$ and occupants $\Phi_{\text{int,OC}}$ [237]:

$$\Phi_{\text{int}} = \Phi_{\text{int,A}} + \Phi_{\text{int,OC}} = 2A_{\text{f}} + 1.5A_{\text{f}} = 3.5A_{\text{f}} \tag{B.22}$$

- the part of the heat flow rate from internal heat sources going to the air node Φ_{int} Φ_{ia} [W] [237]

$$\Phi_{\text{ia}} = \frac{1}{2}\Phi_{\text{int}} \tag{B.23}$$

Given these input parameters, the Crank-Nicholson scheme is defined in [237] is applied. We seek to find the temperature of the internal air node θ_{air} [^{o}C] and of the building mass $\theta_{\text{m,t}}$ at each time step given the heating or cooling power Φ_{HC} (positive for heating and negative for cooling), the external air temperature θ_{e} [^{o}C] and the heat flow rates from solar heat sources Φ_{sol}.

The air node temperature θ_{air} is given as

$$\theta_{\text{air}} = \frac{H_{\text{tr,is}}\theta_{\text{s}} + H_{\text{ve}}\theta_{\text{sup}} + \Phi_{\text{ia}} + \Phi_{\text{HC}}}{H_{\text{tr,is}} + H_{\text{ve}}} \tag{B.24}$$

Where the surface node temperature θ_{s} is defined as:

$$\theta_{\text{s}} = \frac{H_{\text{tr,ms}}\theta_{\text{m}} + \Phi_{\text{st}} + H_{\text{tr,w}}\theta_{\text{e}} + H_{\text{tr,1}}\left(\theta_{\text{sup}} + \frac{\Phi_{\text{ia}}+\Phi_{\text{HC}}}{H_{\text{ve}}}\right)}{H_{\text{tr,ms}} + H_{\text{tr,w}} + H_{\text{tr,1}}} \tag{B.25}$$

The average temperature over the hour of the building mass θ_{m}:

$$\theta_{\text{m}} = \frac{1}{2}(\theta_{\text{m,t-1}} + \theta_{\text{m,t}}) \tag{B.26}$$

$$\theta_{\text{m,t}} = \frac{\theta_{\text{m,t-1}}\left(\frac{C_{\text{m}}}{\tau} + \frac{1}{2}(H_3 + H_{\text{tr,em}})\right) + \Phi_{\text{mtot}}}{\frac{C_{\text{m}}}{\tau} + \frac{1}{2}(H_{\text{tr,3}} + H_{\text{tr,em}})} \tag{B.27}$$

$$\Phi_{\text{mtot}} = \Phi_{\text{m}} + H_{\text{tr,em}}\theta_{\text{e}} + H_{\text{tr,3}}\frac{\Phi_{\text{st}} + H_{\text{tr,w}}\theta_e + H_{\text{tr,1}}\left(\frac{\Phi_{\text{ia}}+\Phi_{\text{HC}}}{H_{\text{ve}}} + \theta_{\text{sup}}\right)}{H_{\text{tr,2}}} \tag{B.28}$$

The part of heat flow rates from internal and solar heat sources going to the internal nodes θ_{s}

$$\Phi_{\text{st}} = \left(1 - \frac{A_{\text{m}}}{A_{\text{t}}} - \frac{H_{\text{tr,w}}}{9.1A_{\text{t}}}\right)\left(\frac{1}{2}\Phi_{\text{int}} + \Phi_{\text{sol}}\right) \tag{B.29}$$

The part of heat flow rates from internal and solar heat sources going to the internal nodes θ_{m}

$$\Phi_{\mathrm{m}} = \frac{A_{\mathrm{m}}}{A_{\mathrm{t}}} \left(\frac{1}{2} \Phi_{\mathrm{int}} + \Phi_{\mathrm{sol}} \right) \tag{B.30}$$

$$\theta_{\mathrm{sup}} = T_{\mathrm{e}} \tag{B.31}$$

We rearrange the equations of this model in order to obtain a linear recursive formulation. We first define some helper variables:

$$A = \frac{C_{\mathrm{m}}}{\tau} + \frac{1}{2}(H_{\mathrm{tr},3} + H_{\mathrm{tr,em}}) \tag{B.32}$$

$$B = 1 - \frac{A_{\mathrm{m}}}{A_{\mathrm{t}}} - \frac{H_{\mathrm{tr,w}}}{9.1 A_{\mathrm{t}}} \tag{B.33}$$

$$C = \frac{B \Phi_{\mathrm{int}}}{2} \tag{B.34}$$

$$D = \frac{A_{\mathrm{m}} \Phi_{\mathrm{int}}}{2 A_{\mathrm{t}}} + \frac{H_{\mathrm{tr},3}}{H_{\mathrm{tr},2}} \left(C + \frac{H_{\mathrm{tr},1} \Phi_{\mathrm{ia}}}{H_{\mathrm{ve}}} \right) \tag{B.35}$$

$$E = H_{\mathrm{tr,em}} + \frac{H_{\mathrm{tr},3}}{H_{\mathrm{tr},2}} (H_{\mathrm{tr,w}} + H_{\mathrm{tr},1}) \tag{B.36}$$

$$H_{\mathrm{tr,ms}} + H_{\mathrm{tr,w}} + H_{\mathrm{tr},1} \tag{B.37}$$

$$G = \frac{1}{F} \left(\frac{H_{\mathrm{tr,ms}} a_{\mathrm{T}}}{2} + C + \frac{H_{\mathrm{tr},1} \Phi_{\mathrm{ia}}}{H_{\mathrm{ve}}} \right) \tag{B.38}$$

$$H = \frac{H_{\mathrm{tr,ms}}}{2F} (1 + b_{\mathrm{T}}) \tag{B.39}$$

$$I = \frac{1}{F} \left(\frac{H_{\mathrm{tr,ms}} c_{\mathrm{T}}}{2} + H_{\mathrm{tr,w}} + H_{\mathrm{tr},1} \right) \tag{B.40}$$

$$J = \frac{1}{F} \left(\frac{H_{\mathrm{tr,ms}} d_{\mathrm{T}}}{2} + B \right) \tag{B.41}$$

$$K = \frac{1}{F}\left(\frac{H_{\mathrm{tr,ms}} e_{\mathrm{T}}}{2} + \frac{H_{\mathrm{tr},1}}{H_{\mathrm{ve}}}\right) \tag{B.42}$$

$$a_{\mathrm{T}} = \frac{D}{A} \tag{B.43}$$

$$b_{\mathrm{T}} = \frac{\left(\frac{C_{\mathrm{m}}}{\tau} + 0.5(H_3 + H_{\mathrm{tr,em}})\right)}{A} \tag{B.44}$$

$$c_{\mathrm{T}} = \frac{E}{A} \tag{B.45}$$

$$d_{\mathrm{T}} = \frac{\frac{A_{\mathrm{m}}}{A_{\mathrm{t}}} + \frac{H_{\mathrm{tr},3} B}{H_{\mathrm{tr},2}}}{A} \tag{B.46}$$

$$e_{\mathrm{T}} = \frac{H_{\mathrm{tr},3} H_{\mathrm{tr},1}}{H_{\mathrm{tr},2} H_{\mathrm{ve}} A} \tag{B.47}$$

$$a_{\mathrm{air}} = \frac{H_{\mathrm{tr,is}} G + \Phi_{\mathrm{ia}}}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} \tag{B.48}$$

$$b_{\mathrm{air}} = \frac{H_{\mathrm{tr,is}} H}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} \tag{B.49}$$

$$c_{\mathrm{air}} = \frac{H_{\mathrm{tr,is}} I + H_{\mathrm{ve}}}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} \tag{B.50}$$

$$d_{\mathrm{air}} = \frac{H_{\mathrm{tr,is}} J}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} \tag{B.51}$$

$$e_{\mathrm{air}} = \frac{H_{\mathrm{tr,is}} K + 1}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} \tag{B.52}$$

$$\xi = \begin{bmatrix} a_{\mathrm{T}} & b_{\mathrm{T}} & c_{\mathrm{T}} & d_{\mathrm{T}} & e_{\mathrm{T}} \\ a_{\mathrm{air}} & b_{\mathrm{air}} & c_{\mathrm{air}} & d_{\mathrm{air}} & e_{\mathrm{air}} \end{bmatrix} \tag{B.53}$$

Rearranging eqs. (B.29), (B.33) and (B.34):

$$\Phi_{\mathrm{st}} = C + B\Phi_{\mathrm{sol}} \tag{B.54}$$

Rearranging eqs. (B.28), (B.30), (B.31) and (B.54):

$$\Phi_{\mathrm{mtot}} = \left(\frac{A_{\mathrm{m}}}{2A_{\mathrm{t}}}\Phi_{\mathrm{int}} + \frac{A_{\mathrm{m}}}{A_{\mathrm{t}}}\Phi_{\mathrm{sol}}\right) + H_{\mathrm{tr,em}}T_{\mathrm{e}} + \frac{H_{\mathrm{tr},3}}{H_{\mathrm{tr},2}}\left(C + B\Phi_{\mathrm{sol}}\right) + \frac{H_{\mathrm{tr},3}}{H_{\mathrm{tr},2}}H_{\mathrm{tr,w}}T_e + \frac{H_{\mathrm{tr},3}}{H_{\mathrm{tr},2}}H_{\mathrm{tr},1}\left(\frac{\Phi_{\mathrm{ia}} + \Phi_{\mathrm{HC}}}{H_{\mathrm{ve}}} + (T_{\mathrm{e}})\right) \tag{B.55}$$

$$\Phi_{\mathrm{mtot}} = \frac{A_{\mathrm{m}}\Phi_{\mathrm{int}}}{2A_{\mathrm{t}}} + \frac{A_{\mathrm{m}}}{A_{\mathrm{t}}}\Phi_{\mathrm{sol}} + H_{\mathrm{tr,em}}T_{\mathrm{e}} + \frac{H_{\mathrm{tr},3}C}{H_{\mathrm{tr},2}} + \frac{H_{\mathrm{tr},3}B}{H_{\mathrm{tr},2}}\Phi_{\mathrm{sol}} + \frac{H_{\mathrm{tr},3}H_{\mathrm{tr,w}}}{H_{\mathrm{tr},2}}T_e + \frac{H_{\mathrm{tr},3}H_{\mathrm{tr},1}\Phi_{\mathrm{ia}}}{H_{\mathrm{tr},2}H_{\mathrm{ve}}} + \frac{H_{\mathrm{tr},3}H_{\mathrm{tr},1}}{H_{\mathrm{tr},2}H_{\mathrm{ve}}}\Phi_{\mathrm{HC}} + \frac{H_{\mathrm{tr},3}H_{\mathrm{tr},1}}{H_{\mathrm{tr},2}}T_{\mathrm{e}} \tag{B.56}$$

$$\Phi_{\mathrm{mtot}} = \left(\frac{A_{\mathrm{m}}\Phi_{\mathrm{int}}}{2A_{\mathrm{t}}} + \frac{H_{\mathrm{tr},3}}{H_{\mathrm{tr},2}}\left(C + \frac{H_{\mathrm{tr},1}\Phi_{\mathrm{ia}}}{H_{\mathrm{ve}}}\right)\right) + \left(\frac{A_{\mathrm{m}}}{A_{\mathrm{t}}} + \frac{H_{\mathrm{tr},3}B}{H_{\mathrm{tr},2}}\right)\Phi_{\mathrm{sol}} + \left(H_{\mathrm{tr,em}} + \frac{H_{\mathrm{tr},3}}{H_{\mathrm{tr},2}}(H_{\mathrm{tr,w}} + H_{\mathrm{tr},1})\right)T_{\mathrm{e}} + \frac{H_{\mathrm{tr},3}H_{\mathrm{tr},1}}{H_{\mathrm{tr},2}H_{\mathrm{ve}}}\Phi_{\mathrm{HC}} \tag{B.57}$$

Rearranging eqs. (B.35), (B.36) and (B.57):

$$\Phi_{\mathrm{mtot}} = D + \left(\frac{A_{\mathrm{m}}}{A_{\mathrm{t}}} + \frac{H_{\mathrm{tr},3}B}{H_{\mathrm{tr},2}}\right)\Phi_{\mathrm{sol}} + ET_{\mathrm{e}} + \frac{H_{\mathrm{tr},3}H_{\mathrm{tr},1}}{H_{\mathrm{tr},2}H_{\mathrm{ve}}}\Phi_{\mathrm{HC}} \tag{B.58}$$

From eqs. (B.27) and (B.32)

$$T_{\mathrm{m,t}} = \frac{\left(\frac{C_{\mathrm{m}}}{\tau} + 0.5(H_3 + H_{\mathrm{tr,em}})\right)}{A}T_{\mathrm{m,t\text{-}1}} + \frac{\Phi_{\mathrm{mtot}}}{A} \tag{B.59}$$

From eqs. (B.58) and (B.59)

$$T_{\mathrm{m,t}} = \frac{D}{A} + \frac{\frac{C_{\mathrm{m}}}{\tau} + 0.5(H_3 + H_{\mathrm{tr,em}})}{A}T_{\mathrm{m,t\text{-}1}} + \frac{E}{A}T_{\mathrm{e}} + \frac{\frac{A_{\mathrm{m}}}{A_{\mathrm{t}}} + \frac{H_{\mathrm{tr},3}B}{H_{\mathrm{tr},2}}}{A}\Phi_{\mathrm{sol}} + \frac{H_{\mathrm{tr},3}H_{\mathrm{tr},1}}{H_{\mathrm{tr},2}H_{\mathrm{ve}}A}\Phi_{\mathrm{HC}} \tag{B.60}$$

From eqs. (B.43) to (B.47) and (B.60):

$$T_{\mathrm{m,t}} = a_{\mathrm{T}} + b_{\mathrm{T}}T_{\mathrm{m,t\text{-}1}} + c_{\mathrm{T}}T_{\mathrm{e}} + d_{\mathrm{T}}\Phi_{\mathrm{sol}} + e_{\mathrm{T}}\Phi_{\mathrm{HC}} \tag{B.61}$$

From eqs. (B.26) and (B.61):

$$T_{\mathrm{m}} = \frac{1}{2}\left(T_{\mathrm{m,t\text{-}1}} + a_{\mathrm{T}} + b_{\mathrm{T}}T_{\mathrm{m,t\text{-}1}} + c_{\mathrm{T}}T_{\mathrm{e}} + d_{\mathrm{T}}\Phi_{\mathrm{sol}} + e_{\mathrm{T}}\Phi_{\mathrm{HC}}\right) \tag{B.62}$$

$$T_{\mathrm{m}} = \frac{a_{\mathrm{T}}}{2} + \frac{1+b_{\mathrm{T}}}{2} T_{\mathrm{m,t\text{-}1}} + \frac{c_{\mathrm{T}}}{2} T_{\mathrm{e}} + \frac{d_{\mathrm{T}}}{2} \Phi_{\mathrm{sol}} + \frac{e_{\mathrm{T}}}{2} \Phi_{\mathrm{HC}} \tag{B.63}$$

From eqs. (B.25), (B.31), (B.37), (B.54) and (B.63)

$$\begin{aligned} T_{\mathrm{s}} = \frac{H_{\mathrm{tr,ms}} a_{\mathrm{T}}}{2F} + \frac{H_{\mathrm{tr,ms}}}{F} \frac{1+b_{\mathrm{T}}}{2} T_{\mathrm{m,t\text{-}1}} + \frac{H_{\mathrm{tr,ms}} c_{\mathrm{T}}}{2F} T_{\mathrm{e}} + \frac{H_{\mathrm{tr,ms}} d_{\mathrm{T}}}{2F} \Phi_{\mathrm{sol}} + \\ \frac{H_{\mathrm{tr,ms}} e_{\mathrm{T}}}{2F} \Phi_{\mathrm{HC}} + \frac{C}{F} + \frac{B}{F} \Phi_{\mathrm{sol}} + \frac{H_{\mathrm{tr,w}}}{F} T_{\mathrm{e}} + \frac{H_{\mathrm{tr,1}}}{F} T_{\mathrm{e}} + \frac{H_{\mathrm{tr,1}} \Phi_{\mathrm{ia}}}{F H_{\mathrm{ve}}} + \frac{H_{\mathrm{tr,1}}}{F H_{\mathrm{ve}}} \Phi_{\mathrm{HC}} \end{aligned} \tag{B.64}$$

$$\begin{aligned} T_{\mathrm{s}} = \frac{1}{F} \left(\frac{H_{\mathrm{tr,ms}} a_{\mathrm{T}}}{2} + C + \frac{H_{\mathrm{tr,1}} \Phi_{\mathrm{ia}}}{H_{\mathrm{ve}}} \right) + \frac{H_{\mathrm{tr,ms}}}{2F} (1+b_{\mathrm{T}}) T_{\mathrm{m,t\text{-}1}} + \\ \frac{1}{F} \left(\frac{H_{\mathrm{tr,ms}} c_{\mathrm{T}}}{2} + H_{\mathrm{tr,w}} + H_{\mathrm{tr,1}} \right) T_{\mathrm{e}} + \\ \frac{1}{F} \left(\frac{H_{\mathrm{tr,ms}} d_{\mathrm{T}}}{2} + B \right) \Phi_{\mathrm{sol}} + \frac{1}{F} \left(\frac{H_{\mathrm{tr,ms}} e_{\mathrm{T}}}{2} + \frac{H_{\mathrm{tr,1}}}{H_{\mathrm{ve}}} \right) \Phi_{\mathrm{HC}} \end{aligned} \tag{B.65}$$

From eqs. (B.38) to (B.42) and (B.65):

$$T_{\mathrm{s}} = G + H T_{\mathrm{m,t\text{-}1}} + I T_{\mathrm{e}} + J \Phi_{\mathrm{sol}} + K \Phi_{\mathrm{HC}} \tag{B.66}$$

From eqs. (B.24), (B.31) and (B.66):

$$\begin{aligned} T_{\mathrm{air}} = \frac{H_{\mathrm{tr,is}} G}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} + \frac{H_{\mathrm{tr,is}} H}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} T_{\mathrm{m,t\text{-}1}} + \frac{H_{\mathrm{tr,is}} I}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} T_{\mathrm{e}} + \frac{H_{\mathrm{tr,is}} J}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} \Phi_{\mathrm{sol}} + \\ \frac{H_{\mathrm{tr,is}} K}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} \Phi_{\mathrm{HC}} + \frac{H_{\mathrm{ve}}}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} T_{\mathrm{e}} + \frac{\Phi_{\mathrm{ia}}}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} + \frac{1}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} \Phi_{\mathrm{HC}} \end{aligned} \tag{B.67}$$

$$\begin{aligned} T_{\mathrm{air}} = \frac{H_{\mathrm{tr,is}} G + \Phi_{\mathrm{ia}}}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} + \frac{H_{\mathrm{tr,is}} H}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} T_{\mathrm{m,t\text{-}1}} + \frac{H_{\mathrm{tr,is}} I + H_{\mathrm{ve}}}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} T_{\mathrm{e}} + \frac{H_{\mathrm{tr,is}} J}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} \Phi_{\mathrm{sol}} + \\ \frac{H_{\mathrm{tr,is}} K + 1}{H_{\mathrm{tr,is}} + H_{\mathrm{ve}}} \Phi_{\mathrm{HC}} \end{aligned} \tag{B.68}$$

From eqs. (B.48) to (B.52) and (B.68):

$$T_{\mathrm{air}} = a_{\mathrm{air}} + b_{\mathrm{air}} T_{\mathrm{m,t\text{-}1}} + c_{\mathrm{air}} T_{\mathrm{e}} + d_{\mathrm{air}} \Phi_{\mathrm{sol}} + e_{\mathrm{air}} \Phi_{\mathrm{HC}} \tag{B.69}$$

Note that the notation from [237] was used in this appendix. In this paper, $T_{\mathrm{air},i}^{t+1} \leftarrow T_{\mathrm{air}}$, $T_{\mathrm{m},i}^{t+1} \leftarrow T_{\mathrm{m,t}}$, $T_{\mathrm{m},i}^{t} \leftarrow T_{\mathrm{m,t\text{-}1}}$, $T_{\mathrm{e}}^{t} \leftarrow T_{\mathrm{e}}$, $\Phi^{t} \leftarrow \Phi_{\mathrm{sol}}$, $h_i^t \leftarrow \frac{\Phi_{\mathrm{HC}}}{COP}$, such that from eqs. (B.53), (B.61) and (B.69):

$$\begin{bmatrix} T_{\mathrm{m},i}^{t+1} \\ T_{\mathrm{air},i}^{t+1} \end{bmatrix} = \xi \left[1, T_{\mathrm{m},i}^{t}, \mathrm{T}_{\mathrm{e}}^{t}, \Phi^{t}, COP \times h_i^t \right]^{\intercal} \tag{B.70}$$

This is equivalent to eq. (3.15).

C

Environment supplementary material

As large action spaces can compound the curse of dimensionality and waste exploration resources [298], another aggregated unified action space is also available in the benchmarking environment, where the real power battery, heating and household consumption actions are determined by an aggregated action a^t_{combined}.

C.1 Aggregated action

At each time step, the decision variables in Section 3.1 controlling the flows in and out of the battery $b^t_{\text{in},i}$ and $b^t_{\text{out},i}$, the electric heating consumption h^t_i and the prosumer consumption c^t_i for household i are therefore synthesised into a single variable $a^t_{\text{combined}} \in [0, 1]$ controlling the use of available local flexibility. Figure C.1 shows how consumption (for domestic loads and heat), imports and storage change with a^t_{combined}.

At each step, the fixed requirements for loads, heat and upcoming EV trips are first met. The a^t_{combined} decision then applies to the remaining flexibility. In conditions deemed optimal for energy exports $a_{\text{combined}} = 0$, all initial storage and residual PV generation is exported and flexible loads are delayed. On the other end, a *passive* agent does not utilise its flexibility and uses the *default* action $a_{\text{combined}} = 1$,

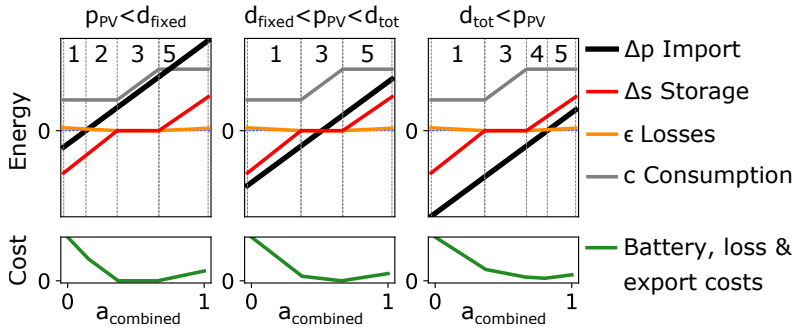


Figure C.1: Combined action a_{combined}. Sections 1-5 denote the trade-off regimes described in Section 3.2.2. At each step, the fixed requirements for loads, heat and upcoming EV trips are first met. The a_{combined} decision then applies to the remaining flexibility, from maximal energy exports (full use of flexibility) at $a_{\text{combined}} = 0$, to maximal energy imports (no use of flexibility) at $a_{\text{combined}} = 1$. d_{tot} and $\mathrm{d}_{\text{fixed}}$ are the sum of household and heating loads with and without their flexible component. If fixed loads cannot be fully met by PV energy, the residual is met by storage and imports (2). If there is additional PV energy after meeting all loads, it can be stored or exported (4).

maximising imports with EVs charged when plugged in and no flexible loads delayed. Intermediate imports trade-offs are mapped on Figure C.1:

1. From exporting all to none of the initial storage E_i^t
2. From meeting fixed loads $\mathrm{d}_{i,\text{fixed}}^t$ with the energy stored to importing the required amount
3. From no to maximum flexible consumption $\mathrm{d}_{i,\text{tot}}^t$
4. From exporting to storing PV energy $\mathrm{p}_{\text{PV},i}^t$ remaining after meeting loads
5. From importing no additional energy to filling up the battery to capacity $\overline{\mathrm{E}}_i$

Costlier actions incurring battery depreciation, losses and export costs are towards either a_{combined}^t extreme, only used in highly beneficial situations (convex local costs function in the lower plot of Figure C.1). For example, it is more cost-efficient to first absorb energy imports by consuming flexible loads, and only use the battery (incurring costs) if imports are large.

C.2 Environment model parameters

Description	Value
Cost of battery depreciation	$C_s = 0.0152$ £/kWh [138][1]
Maximum active charging power	$\overline{b_{in}} = 6.6$ kW [263]
Maximum active discharging power	$\overline{b_{out}} = 7$ kW [299, 300]
Battery capacity	$\overline{E} = 39$ kWh [263]
Minimum state of charge	$\underline{E} = 0.1\overline{E}$
Initial and final state of charge	$E_0 = \overline{E}$
Charging and discharging efficiency	$\eta_{ch} = \eta_{dis} = \sqrt{\eta_{\text{round trip}}}$ [301], $\eta_{\text{round trip}} = 0.87$ [302]
Maximum apparent power	$\overline{S} = 7.0$ kVA
Power factor	$PF = 0.95$ [303]

Table C.1: Electric vehicle battery parameters

Description	Value
Distribution charge on exports	$C_d = 0.01$ £/kWh
Nominal root mean square grid voltage	$V = 415$ [V]
Average resistance between the main grid and the distribution network	$R = 0.084\Omega$
Maximum grid import	$\overline{g} = 6$ kW
Maximum grid export	$\underline{g} = 6$ kW
Penalty on excessive substation imports and exports	$C_s = 0.01$ £/kW
Maximum voltage	$\overline{v} = 1.1$ p.u. [304]
Minimum Voltage	$\underline{v} = 0.94$ p.u. [304]

Table C.2: Network parameters

Description	Value
Initial building temperature	$T_0 = 16^oC$
Set-back temperature	$T_s = 16^oC$
Comfort temperature	$T_c = 20^oC$
Acceptable temperature difference	$\delta T = 2^oC$
Hours when comfort temperature required	7-10 am and 5-10 pm

Table C.3: Heating parameters other than those in Table B.1.

C.3 Home energy data generator (HEDGE) parameters

The parameter values for the GAN profile generator are tabulated in Table C.4.

Parameter	Value

[1] Converting 20 US$/MWh assuming 1 US$ = £0.76

Initial noise	$\epsilon_0 = 1$
End noise	$\epsilon_{\text{end}} = 10^{-4}$
Batch size	$m = 100$
Number of epochs	$n_{\text{epochs}} = 200$
Initial learning rate	$\alpha_0 = 10^{-2}$
End learning rate	$\alpha_{\text{end}} = 10^{-3}$
Number of profiles in generated population	$n_{\text{profiles}} = 50$
Discriminator dropout probability	$p_D = 0.3$
Generator dropout probability	$p_G = 0.15$
Normalised profiles loss weight	$W_1 = 0.1$
Percentile distance loss weight	$W_2 = 100$
EV consumption on motorway	2.25 [kWh/10km] [250]
EV consumption for urban travel	1.62 [kWh/10km] [250]
EV consumption for rural travel	1.36 [kWh/10km] [250]

Table C.4: Generative adversarial network training parameters.

Bibliography

[1] Hannah Ritchie, Max Roser, and Pablo Rosado. Co_2 and greenhouse gas emissions. *Our World in Data*, 2020. https://ourworldindata.org/co2-and-greenhouse-gas-emissions.

[2] International Energy Agency. World Energy Outlook. Technical report, 2022.

[3] BEIS. *Net Zero Strategy: Build Back Greener*. Number October. 2021.

[4] European Commission. An EU Strategy on Heating and Cooling. Technical report, 2016.

[5] International Renewable Energy Agency. Renewable power generation costs in 2020. Technical report, 2021.

[6] Climate Change Committee. The Sixth Carbon Budget: The UK's path to Net Zero. *The Carbon Budget*, (December):34, 2020.

[7] Masson-Delmotte, V. Global Warming of 1.5C. An IPCC Special Report on the impacts of global warming of 1.5C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, 2018.

[8] S. Bose and S. Low. Some Emerging Challenges in Electricity Markets. In *Smart Grid Control*, pages 29–45. Power elec edition, 2019.

[9] J. Vázquez-Canteli and Z. Nagy. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied Energy*, 235(Oct 2018):1072–1089, 2019.

[10] K. Pumphrey, S. Walker, M. Andoni, and V. Robu. Green hope or red herring? Examining consumer perceptions of peer-to-peer energy trading in the United Kingdom. *Energy Research and Social Science*, 68(September 2019):101603, 2020.

[11] Jacob Ostergaard, Charalampos Ziras, Henrik W. Bindner, Jalal Kazempour, Mattia Marinelli, Peter Markussen, Signe Horn Rosted, and Jorgen S. Christensen. Energy security through demand-side flexibility: The case of denmark. *IEEE Power and Energy Magazine*, 19(2):46–55, 2021.

[12] Department for Business Energy and Industrial Strategy. Energy consumption in the UK, 2021.

[13] Vivid Economics and Imperial College London. Accelerated electrification and the GB electricity system, report prepared for Committee on Climate Change. (April):1–79, 2019.

[14] T-O. Léautier. *Imperfect Markets and Imperfect Regulation: An Introduction to the Microeconomics and Political Economy of Power Markets*. MIT Press, 2019.

[15] S. J. Darby. Demand response and smart technology in theory and practice: Customer experiences and system actors. *Energy Policy*, 143(April):111573, 2020.

[16] BloomberNEF. 2019 Battery Price Survey, 2019.

[17] BloomberNEF. 2022 Battery Price Survey, 2012.

[18] Department for Transport Statistics. VEH0101 Licensed vehicles at the end of the quarter by body type, Great Britain from 1994 Q1; also United Kingdom from 2014 Q3, 2020.

[19] National Grid ESO. Winter Outlook Report. Technical Report October, National Grid ESO, 2022.

[20] National Statistics. Digest of UK Energy Statistics (DUKES), 2020.

[21] Brian Drysdale, Jianzhong Wu, and Nick Jenkins. Flexible demand in the gb domestic electricity sector in 2030. *Applied energy*, 139(C):281–290, 2015.

[22] Scott P. Burger, Jesse D. Jenkins, Samuel C. Huntington, and Ignacio J. Perez-Arriaga. Why distributed?: A critical review of the tradeoffs between centralized and decentralized resources. *IEEE power & energy magazine*, 17(2):16–24, 2019.

[23] DNV KEMA. Energy & sustainability. potential for smart electric thermal storage. contributing to a low carbon electricity system, 2016.

[24] Nick Eyre and Pranab Baruah. Uncertainties in future energy demand in UK residential heating. *Energy Policy*, 87:641–653, 2015.

[25] Lesia Marie-Jeanne Mariane Mitridati, Jalal Kazempour, and Pierre Pinson. Design and game-theoretic analysis of community-based market mechanisms in heat and electricity systems. *Omega: The International Journal of Management Science*, 99, 2020.

[26] Ofgem. Enabling the transition to electric vehicles: the regulator's priorities for a green, fair future. Technical report, Ofgem, 2021.

[27] Eoghan McKenna and Murray Thomson. High-resolution stochastic integrated thermal–electrical domestic demand model. *Applied Energy*, 165:445–461, 2016.

[28] Matteo Muratori. Impact of uncoordinated plug-in electric vehicle charging on residential power demand. *Nature Energy*, 3(3):193–201, 2018.

[29] Jesus Lizana, Daniel Friedrich, Renaldi Renaldi, and Ricardo Chacartegui. Energy flexible building through smart demand-side management and latent heat storage. *Applied Energy*, 230(May):471–485, 2018.

[30] Haider Tarish Haider, Ong Hang See, and Wilfried Elmenreich. A review of residential demand response of smart grid. *Renewable and Sustainable Energy Reviews*, 59:166–178, 2016.

[31] Miles Ellingham and Peter Foster. Warm banks help thousands survive cold snap as uk fuel poverty soars. *Financial Times*.

[32] The European Parliament and The Council of the European. Directive (EU) 2019/944 of the European Parliament and of the Council of 5 June 2019 on common rules for the internal market for electricity and amending Directive 2012/27/EU, 2019.

[33] T. Sousa, T. Soares, P. Pinson, F. Moret, T. Baroche, and E. Sorin. Peer-to-peer and community-based markets: A comprehensive review. *Renewable and Sustainable Energy Reviews*, 104:367–378, 2019.

[34] Charles River Associates. An assessment of the economic value of demand-side participation in the Balancing Mechanism and an evaluation of options to improve access, 2017.

[35] T. Chen and W. Su. Indirect Customer-to-Customer Energy Trading with Reinforcement Learning. *IEEE Transactions on Smart Grid*, 10(4):4338–4348, 2019.

[36] D. Bugden and R. Stedman. A synthetic view of acceptance and engagement with smart meters in the United States. *Energy Research and Social Science*, 47(January 2018):137–145, 2019.

[37] F. Moret and P. Pinson. Energy Collectives: A Community and Fairness Based Approach to Future Electricity Markets. *IEEE Transactions on Power Systems*, 34(5):3994–4004, 2019.

[38] Chase, A. Realising the Potential of Demand-Side Response to 2025: A focus on Small Energy Users Rapid Evidence Assessment report, 2017.

[39] UK Office for National Statistics. Families and households in the UK: 2020. Technical report, Office for National Statistics, 2021.

[40] Samuel Gyamfi, Susan Krumdieck, and Tania Urmee. Residential peak electricity demand response—highlights of some behavioural issues. *Renewable & sustainable energy reviews*, 25:71–77, 2013.

[41] J. Torriti, M. G Hassan, and M. Leach. Demand response experience in europe: Policies, programmes and implementation. *Energy (Oxford)*, 35(4):1575–1583, 2010.

[42] Sarah J. Darby. Smart and sustainable, fast and slow. *Eceee Summer Study Proceedings*, 2019-June:939–948, 2019.

[43] Local Energy Oxfordshire. Project leo final report: A digest of key learnings. Technical report, Local Energy Oxfordshire, February 2023.

[44] Charlotte Reypens and Jonathan Bone. Project updates - where next for peer-to-peer energy exchange?, 2021.

[45] Elexon. P379 Impact Assessment. Technical Report March, Elexon, 2020.

[46] Fergal Egan. The Dingle Electrification Project: Sharing Learnings from the Peer-to-Peer Energy Trading Objective. Technical Report December, ESB Networks, 2020.

[47] David Livingston, Varun Sivaram, Madison Freeman, and Maximilian Fiege. Applying blockchain technology to electric power systems. Technical report, Council on Foreign Relations, 2018.

[48] European Investment Bank. The eib climate survey 2019-2020: How citizens are confronting the climate crisis and what actions they expect from policymakers and businesses. Technical report, European Investment Bank, 2020.

[49] Ipsos. Earth day 2022: Global attitudes on climate change, April 2022. Version 1.

[50] Cristina Rottondi and Giacomo Verticale. A privacy-friendly gaming framework in smart electricity and water grids. 5:14221–14233, 2017.

[51] C. Laughman, Kwangduk Lee, R. Cox, S. Shaw, S. Leeb, L. Norford, and P. Armstrong. Power signature analysis. *IEEE Power and Energy Magazine*, 1(2):56–63, 2003.

[52] Jiyun Yao. *Cybersecurity of Demand Side Management in the Smart Electricity Grid: Privacy Protection, Battery Capacity Sharing and Power Grid under Attack*. PhD thesis, 2017. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2021-05-20.

[53] Kate Crawford. *The Atlas of AI*. Yale University Press, apr 2021.

[54] European Parliament and Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016. Official Journal of the European Union L 119, 4.5.2016, p. 1-88.

[55] Daniel J. Solove. *Nothing to Hide: The False Tradeoff between Privacy and Security*. Yale University Press, New Haven, CT, 2013.

[56] European Union Agency for Fundamental Rights. *Fundamental Rights in the Digital Age*. Publications Office of the European Union, Luxembourg, 2018.

[57] Data Protection Commission. *Annual Report 2018*. Data Protection Commission, Dublin, Ireland, 2019.

[58] O Sachs. Field Evaluation of Programmable Thermostats, 2012.

[59] Element Energy Limited, Energy Systems Catapult, Cenex, Nissan Technical Centre Europe, Western Power Distribution, and National Grid ESO. Vehicle to Grid Britain. 2019.

[60] Marvin Gleue, Jens Unterberg, Andreas Löschel, and Philipp Grünewald. Does demand-side flexibility reduce emissions? Exploring the social acceptability of demand management in Germany and Great Britain. *Energy Research and Social Science*, 82(September):102290, 2021.

[61] Xiaoping He and David Reiner. Why Do More British Consumers Not Switch Energy Suppliers? The Role of Individual Attitudes (EPRG Working Paper). (September 2015):1–34, 2015.

[62] Elisabeth Dütschke and Alexandra-Gwyn Paetz. Dynamic electricity pricing—which programs do consumers prefer? *Energy Policy*, 59:226–234, 2013.

[63] Sarah Darby, Jessica Strömbäck, and Mike Wilks. Potential carbon impacts of smart grid development in six european countries. *Energy efficiency*, 6(4):725–739, 2013.

[64] Christian Schlereth, Bernd Skiera, and Fabian Schulz. Why do consumers prefer static instead of dynamic pricing plans? an empirical study for a better understanding of the low preferences for time-variant pricing plans. *European journal of operational research*, 269(3):1165–1179, 2018.

[65] H. H. Happ. Optimal power dispatch — a comprehensive survey. *IEEE Transactions on Power Apparatus and Systems*, 96(3):841–854, 1977.

[66] Stephen Boyd. *Convex optimization theory*, volume 25. 2009.

[67] T. Morstyn, A. Teytelboym, C. Hepburn, and M. McCulloch. Integrating P2P Energy Trading with Probabilistic Distribution Locational Marginal Pricing. *IEEE Transactions on Smart Grid*, 11(4):3095–3106, 2020.

[68] Lixing Chen and Jie Xu. Socially trusted collaborative edge computing in ultra dense networks. *2017 2nd ACM/IEEE Symposium on Edge Computing, SEC 2017*, 2017.

[69] Viktorija Dudjak, Diana Neves, Tarek Alskaif, Shafi Khadem, Alejandro Pena-bello, Pietro Saggese, Benjamin Bowler, Merlinda Andoni, Marina Bertolini, Yue Zhou, Blanche Lormeteau, Mustafa A Mustafa, Yingjie Wang, Christina Francis, Fairouz Zobiri, David Parra, and Antonios Papaemmanouil. Impact of local energy markets integration in power systems layer : A comprehensive review. *Applied Energy*, 301(March):117434, 2021.

[70] C IRENA. Innovation landscape for a renewable-powered future: solutions to integrate variable renewables, 2019.

[71] Vítor Monteiro, Henrique Gonçalves, and João L. Afonso. Impact of electric vehicles on power quality in a smart grid context. In *11th International Conference on Electrical Power Quality and Utilisation*, pages 1–6, 2011.

[72] B IRENA. Innovation landscape brief: Innovative ancillary services, 2019.

[73] Thomas Morstyn, Niall Farrell, Sarah J. Darby, and Malcolm D. McCulloch. Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants. *Nature Energy*, 3(2):94–101, 2018.

[74] Oliver Mihatsch and Ralph Neuneier. Risk-Sensitive Reinforcement Learning. *Machine Learning*, pages 267–290, 2002.

[75] Peter Geibel and Fritz Wysotzki. Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research*, 24:81–108, 2005.

[76] L. Han, T. Morstyn, and M. McCulloch. Incentivizing Prosumer Coalitions With Energy Management Using Cooperative Game Theory. *IEEE Transactions on Power Systems*, 34(1):303–313, 2019.

[77] United Nations. Resolution adopted by the General Assembly on 25 September 2015: Transforming our world: the 2030 Agenda for Sustainable Development. Technical report, 2015.

[78] Ofgem. Targeted charging review: minded to decision and draft impact assessment. (December):1–24, 2019.

[79] Hieu Trung Nguyen, Swathi Battula, Rohit Reddy Takkala, Zhaoyu Wang, and Leigh Tesfatsion. An integrated transmission and distribution test system for evaluation of transactive energy designs. *Applied Energy*, 240(August 2018):666–679, 2019.

[80] taxonomy, n. In *Oxford English Dictionary*. Oxford University Press, 2021.

[81] Akintonde O. Abbas and Badrul H. Chowdhury. Using customer-side resources for market-based transmission and distribution level grid services – A review. *International Journal of Electrical Power and Energy Systems*, 125(May 2020):106480, 2021.

[82] Cherrelle Eid, Paul Codani, Yannick Perez, Javier Reneses, and Rudi Hakvoort. Managing electric flexibility from Distributed Energy Resources: A review of incentives for market design. *Renewable and Sustainable Energy Reviews*, 64:237–247, 2016.

[83] Omid Abrishambaf, Fernando Lezama, Pedro Faria, and Zita Vale. Towards transactive energy systems: An analysis on current trends. *Energy Strategy Reviews*, 26:100418, 2019.

[84] Pierluigi Siano. Demand response and smart grids - A survey. *Renewable and Sustainable Energy Reviews*, 30:461–478, 2014.

[85] P. Kohlhepp, H. Harb, H. Wolisz, S. Waczowicz, D. Müller, and V. Hagenmeyer. Large-scale grid integration of residential thermal energy storages as demand-side flexibility resource: A review of international field studies. *Renewable and Sustainable Energy Reviews*, 101(December 2018):527–547, 2019.

[86] Qin Zhang and Juan Li. Demand response in electricity markets: A review. *9th International Conference on the European Energy Market, EEM 12*, 2012.

[87] Ijaz Hussain, Sajjad Mohsin, Abdul Basit, Zahoor Ali Khan, Umar Qasim, and Nadeem Javaid. A review on demand response: Pricing, optimization, and appliance scheduling. *Procedia Computer Science*, 52(1):843–850, 2015.

[88] Lexuan Meng, Eleonora Riva Sanseverino, Adriana Luna, Tomislav Dragicevic, Juan C. Vasquez, and Josep M. Guerrero. Microgrid supervisory controllers and energy management systems: A literature review. *Renewable and Sustainable Energy Reviews*, 60:1263–1273, 2016.

[89] T. Morstyn, B. Hredzak, and V. Agelidis. Control Strategies for Microgrids with Distributed Energy Storage Systems: An Overview. *IEEE Transactions on Smart Grid*, 9(4):3652–3666, 2018.

[90] Usman Bashir Tayab, Mohd Azrik Bin Roslan, Leong Jenn Hwai, and Muhammad Kashif. A review of droop control techniques for microgrid. *Renewable and Sustainable Energy Reviews*, 76(March):717–727, 2017.

[91] F. Bandeiras, E. Pinheiro, M. Gomes, P. Coelho, and J. Fernandes. Review of the cooperation and operation of microgrid clusters. *Renewable and Sustainable Energy Reviews*, 133(August):110311, 2020.

[92] Mahdi Behrangrad. A review of demand side management business models in the electricity market. *Renewable and Sustainable Energy Reviews*, 47:270–283, 2015.

[93] Theodor Borsche and Goran Andersson. A review of demand response business cases. *IEEE PES Innovative Smart Grid Technologies Conference Europe*, 2015-January(January):1–6, 2015.

[94] Junjie Hu, Guangya Yang, Koen Kok, Yusheng Xue, and Henrik W. Bindner. Transactive control: a framework for operating power systems characterized by high penetration of distributed energy resources. *Journal of Modern Power Systems and Clean Energy*, 5(3):451–464, 2017.

[95] Muhammad F. Zia, Mohamed Benbouzid, Elhoussin Elbouchikhi, S. M. Muyeen, Kuaanan Techato, and Josep M. Guerrero. Microgrid transactive energy: Review, architectures, distributed ledger technologies, and market analysis. *IEEE Access*, 8:19410–19432, 2020.

[96] W. Tushar, C. Yuen, T. Saha, T. Morstyn, A. Chapman, M. Alam, S. Hanif, and V. Poor. Peer-to-peer energy systems for connected communities: A review of recent advances and emerging challenges. *Applied Energy*, 282(PA):116131, 2021.

[97] R. Machlev, N. Zargari, N. R. Chowdhury, J. Belikov, and Y. Levron. A review of optimal control methods for energy storage systems - energy trading, energy balancing and electric vehicles. *Journal of Energy Storage*, 32(July):101787, 2020.

[98] C. Schellenberg, J. Lohan, and L. Dimache. Comparison of metaheuristic optimisation methods for grid-edge technology that leverages heat pumps and thermal energy storage. *Renewable and Sustainable Energy Reviews*, 131(June):109966, 2020.

[99] Pierluigi Siano, Giuseppe De Marco, Alejandro Rolan, and Vincenzo Loia. A Survey and Evaluation of the Potentials of Distributed Ledger Technology for Peer-to-Peer Transactive Energy Exchanges in Local Energy Markets. *IEEE Systems Journal*, 13(3):3454–3466, 2019.

[100] Tam T. Mai, Phuong H. Nguyen, Quoc Tuan Tran, Alessia Cagnano, Giovanni De Carne, Yassine Amirat, Anh Tuan Le, and Enrico De Tuglie. An overview of grid-edge control with the digital transformation. *Electrical Engineering*, 103(4):1989–2007, 2021.

[101] Yaser Tohidi, Mana Farrokhseresht, and Madeleine Gibescu. A review on coordination schemes between local and central electricity markets. *International Conference on the European Energy Market, EEM*, 2018-June, 2018.

[102] J. Guerrero. Towards a transactive energy system for integration of distributed energy resources: Home energy management, distributed optimal power flow, and peer-to-peer energy trading. *Renewable & sustainable energy reviews*, 132, 2020.

[103] Elsevier B.V. Scopus, 2020.

[104] Peter J. Hall and Euan J. Bain. Energy-storage technologies and electricity generation. *Energy Policy*, 36(12):4352–4355, 2008.

[105] Nandha Kumar Kandasamy, King Tseng, and Bh Soong. A virtual storage capacity using demand response management to overcome intermittency of solar pv generation. *IET Renewable Power Generation*, 11, 09 2017.

[106] Benedikt Römer, Philipp Reichhart, Johann Kranz, and Arnold Picot. The role of smart metering and decentralized electricity storage for smart grids: The importance of positive externalities. *Energy policy*, 50:486–495, 2012.

[107] Jeremy Bentham. *An introduction to the principles of morals and legislation*. Clarendon Press, Oxford, 1879.

[108] Nigar Hashimzade, Gareth Myles, and John Black. utility function, 2017.

[109] Robert Wilson. Architecture of power markets. *Econometrica*, 70(4):1299–1340, 2002.

[110] M. Parry. *Climate change 2007: impacts, adaptation and vulnerability*. Published for the Intergovernmental Panel on Climate Change [by] Cambridge University Press, Cambridge, 2007.

[111] M. Wooldridge. *Intelligent Agents: The Key Concepts*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

[112] Rehan Fazal, Jignesh Solanki, and Sarika Khushalani Solanki. Demand response using multi-agent system. *2012 North American Power Symposium, NAPS 2012*, 2012.

[113] Mohsen Khorasany, Yateendra Mishra, and Gerard Ledwich. A Decentralized Bilateral Energy Trading System for Peer-to-Peer Electricity Markets. *IEEE Transactions on Industrial Electronics*, 67(6):4646–4657, 2020.

[114] Marie-Louise Arlt, David P. Chassin, and L. Lynne Kiesling. Opening up transactive systems: Introducing tess and specification in a field deployment. *Energies*, 14(13), 2021.

[115] Koen Kok and Steve Widergren. A Society of Devices: Integrating Intelligent Distributed Resources with Transactive Energy. *IEEE Power and Energy Magazine*, 14(3):34–45, 2016.

[116] B.J. Claessens, S. Vandael, F. Ruelens, K. De Craemer, and B. Beusen. Peak shaving of a heterogeneous cluster of residential flexibility carriers using reinforcement learning. In *IEEE PES ISGT Europe 2013*, pages 1–5, 2013.

[117] M. G. Vayá, L. B. Roselló, and G. Andersson. Optimal bidding of plug-in electric vehicles in a market-based control setup. *Proceedings - 2014 Power Systems Computation Conference, PSCC 2014*, 2014.

[118] Xiaoshun Zhang, Tao Bao, Tao Yu, Bo Yang, and Chuanjia Han. Deep transfer q-learning with virtual leader-follower for supply-demand stackelberg game of smart grid. *Energy*, 133:348–365, 2017.

[119] D. Dauer, C. M. Flath, P. Ströhle, and C. Weinhardt. Market-based EV charging coordination. *Proceedings - 2013 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2013*, 2:102–107, 2013.

[120] B. Kim, Y. Zhang, M. Van Der Schaar, and J. Lee. Dynamic Pricing and Energy Consumption Scheduling With Reinforcement Learning. *IEEE Transactions on Smart Grid*, 7(5):2187–2198, 2016.

[121] Jip Kim and Yury Dvorkin. A P2P-dominant Distribution System Architecture. *IEEE transactions on power systems*, 35:2716–2725, 2019.

[122] Jaysson Guerrero, Archie C. Chapman, and Gregor Verbic. Decentralized P2P Energy Trading under Network Constraints in a Low-Voltage Network. *IEEE Transactions on Smart Grid*, pages 1–10, 2018.

[123] Zhong Zhang, Ran Li, and Furong Li. A Novel Peer-to-Peer Local Electricity Market for Joint Trading of Energy and Uncertainty. *IEEE Transactions on Smart Grid*, 11(2):1205–1215, 2020.

[124] Maria Luisa Di Silvestre, Pierluigi Gallo, Mariano Giuseppe Ippolito, Eleonora Riva Sanseverino, and Gaetano Zizzo. A technical approach to the energy blockchain in microgrids. *IEEE Transactions on Industrial Informatics*, 14(11):4792–4803, 2018.

[125] T. Morstyn, A. Teytelboym, and M. McCulloch. Designing decentralized markets for distribution system flexibility. *IEEE Transactions on Power Systems*, 34(3):1–12, 2019.

[126] Farrokh Rahimi, Ali Ipakchi, and Fred Fletcher. The Changing Electrical Landscape: End-to-End Power System Operation Under the Transactive Energy Paradigm. *IEEE Power and Energy Magazine*, 14(3):52–62, 2016.

[127] I. Dusparic. Multi-agent residential demand response based on load forecasting. *2013 1st IEEE Conference on Technologies for Sustainability, SusTech 2013*, pages 90–96, 2013.

[128] I. Dusparic. Maximizing renewable energy use with decentralized residential demand response. *2015 IEEE 1st International Smart Cities Conference, ISC2 2015*, 2015.

[129] L. A. Hurtado. Enabling Cooperative Behavior for Building Demand Response Based on Extended Joint Action Learning. *IEEE Transactions on Industrial Informatics*, 14(1):127–136, 2018.

[130] W. Tushar, T. Saha, C. Yuen, T. Morstyn, Nahid-Al-Masood, H. Poor, and R. Bean. Grid Influenced Peer-to-Peer Energy Trading. *IEEE Transactions on Smart Grid*, 11(2):1407–1418, 2020.

[131] Liliane Ableitner. Quartierstrom. Implementation of a real world prosumer centric local energy market in Walenstadt, Switzerland. 2019.

[132] Wayes Tushar. A motivational game-theoretic approach for peer-to-peer energy trading in the smart grid. *Applied Energy*, 243(November 2018):10–20, 2019.

[133] B. P. Hayes, S. Thakur, and J. G. Breslin. Co-simulation of electricity distribution networks and peer to peer energy trading platforms. *International Journal of Electrical Power and Energy Systems*, 115(May 2019):105419, 2020.

[134] T. Morstyn, A. Teytelboym, and M. McCulloch. Bilateral contract networks for peer-to-peer energy trading. *IEEE Transactions on Smart Grid*, 10(2):2026–2035, 2019.

[135] A. Taylor. Accelerating Learning in multi-objective systems through Transfer Learning. *Proceedings of the International Joint Conference on Neural Networks*, pages 2298–2305, 2014.

[136] A. Marinescu, I. Dusparic, and S. Clarke. Prediction-based multi-agent reinforcement learning in inherently non-stationary environments. *ACM Transactions on Autonomous and Adaptive Systems*, 12(2), 2017.

[137] A. Pigott, C. Crozier, K. Baker, and Z. Nagy. GridLearn: Multiagent reinforcement learning for grid-aware building energy management. *Electric Power Systems Research*, 213(October 2021):108521, 2022.

[138] T. Morstyn and M. McCulloch. Multiclass Energy Management for Peer-to-Peer Energy Trading Driven by Prosumer Preferences. *IEEE Transactions on Power Systems*, 34(5):4005–4014, 2019.

[139] Amrit Paudel, Kalpesh Chaudhari, Chao Long, and Hoay Beng Gooi. Peer-to-peer energy trading in a prosumer-based community microgrid: A game-theoretic model. *IEEE Transactions on Industrial Electronics*, 66(8):6087–6097, 2019.

[140] J. G. Kim and B. Lee. Automatic P2P energy trading model based on reinforcement learning using long short-term delayed reward. *Energies*, 13(20), 2020.

[141] T. Baroche, P. Pinson, R. Le Goff Latimier, and H. Ben Ahmed. Exogenous Cost Allocation in Peer-to-Peer Electricity Markets. *IEEE Transactions on Power Systems*, 34(4):2553–2564, 2019.

[142] T. Morstyn and M. Mcculloch. Peer-to-Peer Energy Trading. *Analytics for the Sharing Economy: Mathematics, Engineering and Business Perspectives*, (March), 2020.

[143] T. Morstyn, B. Hredzak, R. Aguilera, and V. Agelidis. Model Predictive Control for Distributed Microgrid Battery Energy Storage Systems. *IEEE Transactions on Control Systems Technology*, 26(3):1107–1114, 2018.

[144] Chengda Ji, Pengcheng You, Elijah J Pivo, Yue Shen, Dennice F Gayme, and Enrique Mallada. Optimal Coordination of Distribution System Resources under Uncertainty for Joint Energy and Ancillary Service Market Participation. 2019.

[145] Kai Heussen, Stephan Koch, Andreas Ulbig, and Göran Andersson. Energy storage in power system operation: The power nodes modeling framework. *IEEE PES Innovative Smart Grid Technologies Conference Europe, ISGT Europe*, pages 1–8, 2010.

[146] Philipp Fortenbacher, Johanna L. Mathieu, and Goran Andersson. Modeling and Optimal Operation of Distributed Battery Storage in Low Voltage Grids. *IEEE Transactions on Power Systems*, 32(6):4340–4350, 2017.

[147] J. Cao, C. Crozier, M. McCulloch, and Z. Fan. Optimal design and operation of a low carbon community based multi-energy systems considering EV integration. *IEEE Transactions on Sustainable Energy*, 10(3):1217–1226, 2019.

[148] Ehsan Nasrolahpour, S. Jalal Kazempour, Hamidreza Zareipour, and William D. Rosehart. Strategic sizing of energy storage facilities in electricity markets. *IEEE Transactions on Sustainable Energy*, 7(4):1462–1472, 2016.

[149] D. O'Neill, M. Levorato, A. Goldsmith, and U. Mitra. Residential Demand Response Using Reinforcement Learning. *2010 First IEEE International Conference on Smart Grid Communications*, pages 409–414, 2010.

[150] Michael Stadler, Wolfram Krause, Michael Sonnenschein, and Ute Vogel. The Adaptive Fridge – Comparing different control schemes for enhancing load shifting of electricity demand. *Environmental Protection*, pages 199–206, 2007.

[151] H. Hao, B. M. Sanandaji, K. Poolla, and T. Vincent. Aggregate flexibility of thermostatically controlled loads. *IEEE Transactions on Power Systems*, 30(1):189–198, 2015.

[152] R. Lu and S. H. Hong. Incentive-based demand response for smart grid with reinforcement learning and deep neural network. *Applied Energy*, 236(December 2018):937–949, 2019.

[153] P. H. Babar, M. Zbigniew Hanzelka. The evaluation of agile demand response: An applied methodology. *IEEE Transactions on Smart Grid*, 9(6):6118–6127, 2018.

[154] Sabita Maharjan, Quanyan Zhu, Yan Zhang, Stein Gjessing, and Tamer Başsar. Dependable demand response management in the smart grid: A stackelberg game approach. *IEEE Transactions on Smart Grid*, 4(1):120–132, 2013.

[155] Minghui Zhu. Distributed demand response algorithms against semi-honest adversaries. *IEEE Power and Energy Society General Meeting*, 2014-Octob(October):0–4, 2014.

[156] Bendong Zhao, Huanzhang Lu, Shangfeng Chen, Junliang Liu, and Dongya Wu. Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics*, 28(1):162–169, 2017.

[157] Y. Ye. Deep Reinforcement Learning for Strategic Bidding in Electricity Markets. *IEEE Transactions on Smart Grid*, 11(2):1343–1355, 2020.

[158] Y. Sun, A. Somani, and T. Carroll. Learning based bidding strategy for HVAC systems in double auction retail energy markets. *Proceedings of the American Control Conference*, 2015-July:2912–2917, 2015.

[159] John Black, Nigar Hashimzade, and Gareth Myles. *A Dictionary of Economics*. Oxford University Press, jan 2012.

[160] S. Tindemans, V. Trovato, and G. Strbac. Decentralized Control of Thermostatic Loads for Flexible Demand Response. *IEEE Transactions on Control Systems Technology*, 23(5):1685–1700, 2015.

[161] F.F. Wu and Pravin Varaiya. Coordinated multilateral trades for electric power networks: theory and implementation. *International Journal of Electrical Power and Energy Systems*, 21:75–102, 1999.

[162] Panagiotis Andrianesis and Michael C. Caramanis. Optimal Grid - Distributed Energy Resource Coordination: Distribution Locational Marginal Costs and Hierarchical Decomposition. *2019 57th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2019*, pages 318–325, 2019.

[163] Z. Li, J. Kang, R. Yu, D. Ye, Q. Deng, and Y. Zhang. Consortium blockchain for secure energy trading in industrial internet of things. *IEEE Transactions on Industrial Informatics*, 14(8):3690–3700, 2018. cited By 388.

[164] Robert Herian. Regulating Disruption: Blockchain, GDPR, and Questions of Data Sovereignty. *Journal of Internet Law*, 22(2), 2018.

[165] Andrew L. Goodkind, Benjamin A. Jones, and Robert P. Berrens. Cryptodamages: Monetary value estimates of the air pollution and human health impacts of cryptocurrency mining. *Energy Research and Social Science*, 59(September 2019):101281, 2020.

[166] Eli Hadzhieva. Impact of digitalisation on international tax matters - challenges and remedies. Technical report, European Parliament Policy Department for Economic, Scientific and Quality of Life Policies, 2019.

[167] S. Herbert. *Models of bounded rationality.* MIT Press, Cambridge, Mass. ; London, 1982.

[168] Hongming Yang, Meng Zhang, and Mingyong Lai. Complex dynamics of cournot game with bounded rationality in an oligopolistic electricity market. *Optimization and engineering*, 12(4):559–582, 2011.

[169] Julia Blasch, Massimo Filippini, and Nilkanth Kumar. Boundedly rational consumers, energy and investment literacy, and the display of information on household appliances. *Resource and energy economics*, 56:39–58, 2019.

[170] Emmanuel Farhi and Iván Werning. Monetary policy, bounded rationality, and incomplete markets. *The American economic review*, 109(11):3887–3928, 2019.

[171] Tamas Fleiner, Zsuzsanna Janko, Akihisa Tamura, and Alexander Teytelboym. Trading networks with bilateral contracts. *EAI Endorsed Transactions on Serious Games*, pages 1–39, 2015.

[172] F. L. Lewis, H. Zhang, K. Hengster-Movric, and A. Das. *Cooperative Control of Multi-Agent Systems (Communications and Control Engineering).* Springer, London, U.K., 2014.

[173] Vladimir Dvorkin, Jalal Kazempour, Luis Baringo, and Pierre Pinson. A consensus-admm approach for strategic generation investment in electricity markets. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 780–785, 2018.

[174] Tooraj Jamasb and Michael Pollitt. Incentive regulation of electricity distribution networks: Lessons of experience from Britain. *Energy Policy*, 35(12):6163–6187, 2007.

[175] Sandip Sen, Mahendra Sekaran, and John Hale. Learning to coordinate without sharing information. *Proceedings of the National Conference on Artificial Intelligence*, 1:426–431, 1994.

[176] Jang Won Lee and Du Han Lee. Residential electricity load scheduling for multi-class appliances with Time-of-Use pricing. *2011 IEEE GLOBECOM Workshops, GC Wkshps 2011*, pages 1194–1198, 2011.

[177] J. Cao. Deep Reinforcement Learning Based Energy Storage Arbitrage With Accurate Lithium-ion Battery Degradation Model. *IEEE Transactions on Smart Grid*, 14(8):1–9, 2019.

[178] Hao Wang and Baosen Zhang. Energy Storage Arbitrage in Real-Time Markets via Reinforcement Learning. In *IEEE Power and Energy Society General Meeting*, volume 2018-Augus, pages 1–11, 2018.

[179] Z. Wen, D. O'Neill, and H. Maei. Optimal demand response using device-based reinforcement learning. *IEEE Transactions on Smart Grid*, 6(5):2312–2324, 2015.

[180] K. Dalamagkidis, D. Kolokotsa, K. Kalaitzakis, and G. S. Stavrakakis. Reinforcement learning for energy conservation and comfort in buildings. *Building and Environment*, 42(7):2686–2698, 2007.

[181] Lei Yang, Zoltan Nagy, Philippe Goffin, and Arno Schlueter. Reinforcement learning for optimal control of low exergy buildings. *Applied Energy*, 156:577–586, 2015.

[182] C. Crozier, D. Apostolopoulou, and M. McCulloch. Mitigating the impact of personal vehicle electrification: A power generation perspective. *Energy Policy*, 118(2013):474–481, 2018.

[183] Frank W Geels, Benjamin K Sovacool, Tim Schwanen, and Steve Sorrell. Sociotechnical transitions for deep decarbonization. *Science (New York, N.Y.)*, 357(6357):1242, 2017.

[184] William MacAskill. *Doing good better [electronic resource] : effective altruism and a radical new way to make a difference*. London, 2015.

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

[186] Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *arXiv*, 2018.

[187] Vladimir Dvorkin, Ferdinando Fioretto, Pascal Van Hentenryck, Pierre Pinson, and Jalal Kazempour. Privacy-preserving convex optimization: When differential privacy meets stochastic programming, 2022.

[188] Vincent François Lavet. Contributions to deep reinforcement learning and its applications in smartgrids, 2017.

[189] Subrata Dasgupta. *Computer Science: A Very Short Introduction*. Oxford University Press, mar 2016.

[190] Frederik Ruelens. Residential Demand Response of Thermostatically Controlled Loads Using Batch Reinforcement Learning. *IEEE Transactions on Smart Grid*, 8(5):2149–2159, 2017.

[191] James B. Rawlings, David Q. Mayne, and Moritz M. Diehl. *Model Predictive Control: Theory and Design*. Nob Hill, Madison, WI, USA, second edition, 2017.

[192] Ján Drgoňa, Javier Arroyo, Iago Cupeiro Figueroa, David Blum, Krzysztof Arendt, Donghun Kim, Enric Perarnau Ollé, Juraj Oravec, Michael Wetter, Draguna L Vrabie, et al. All you need to know about model predictive control for buildings. *Annual Reviews in Control*, 50:190–232, 2020.

[193] Jiří Cigler, Dimitrios Gyalistras, Jan Široký, V Tiet, and Lukaš Ferkl. Beyond theory: the challenge of implementing model predictive control in buildings. *Proceedings of 11th Rehva world congress, Clima*, 250, 6 2013.

[194] Richard S. Sutton, Andrew G. Barto, and Ronald J. Williams. Reinforcement Learning is Direct Adaptive Optimal Control. *IEEE Control Systems*, 12(2):19–22, 1992.

[195] Vincent François-lavet, Peter Henderson, Riashat Islam, Marc G Bellemare, Vincent François-lavet, Joelle Pineau, and Marc G Bellemare. An Introduction to Deep Reinforcement Learning. *Foundations and trends in machine learning*, II(3 - 4):1–140, 2018.

[196] Margaret A. Boden. *Artificial Intelligence: A Very Short Introduction*. 2018.

[197] I. Antonopoulos. Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. *Renewable and Sustainable Energy Reviews*, 130(April):109899, 2020.

[198] Christopher John Cornish Hellaby Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, UK, May 1989.

[199] Christopher Watkins and Peter Dayan. Q -learning. *Machine Learning*, 8(3-4):279–292, 1992.

[200] Bob Price and Craig Boutilier. Accelerating reinforcement learning through implicit imitation. *Journal of Artificial Intelligence Research*, 19:569–629, 2003.

[201] D. O Hebb. *The organization of behavior : a neuropsychological theory*. Wiley book in clinical psychology. Wiley, New York, 1949.

[202] L. Buşoniu, R. Babuška, and B. De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 38(2):156–172, 2008.

[203] R. Lowe. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. 2020.

[204] Maximilian Hüttenrauch, Adrian Sosic, and Gerhard Neumann. Guided deep reinforcement learning for swarm systems. *CoRR*, abs/1709.06011, 2017.

[205] Yongcan Cao, Wenwu Yu, Wei Ren, and Guanrong Chen. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial Informatics*, 9(1):427–438, 2013.

[206] L. Matignon, G. Laurent, and N. Le Fort-Piat. Independent reinforcement learners in cooperative Markov games: A survey regarding coordination problems. *Knowledge Engineering Review*, 27(1):1–31, 2012.

[207] Martin Lauer and Martin Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *In Proceedings of the Seventeenth International Conference on Machine Learning*, pages 535–542. Morgan Kaufmann, 2000.

[208] Frans A Oliehoek, Matthijs TJ Spaan, and Nikos Vlassis. Optimal and approximate q-value functions for decentralized pomdps. *Journal of Artificial Intelligence Research*, 32:289–353, 2008.

[209] M. Tan. Multi-Agent Reinforcement Learning : Independent vs . Cooperative Agents. 1993.

[210] V. Mnih. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013.

[211] S. Omidshafiei. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. *34th International Conference on Machine Learning, ICML 2017*, 6:4108–4122, 2017.

[212] J. Foerster. Stabilising experience replay for deep multi-agent reinforcement learning. *34th International Conference on Machine Learning, ICML 2017*, 3:1879–1888, 2017.

[213] F. Charbonnier, Morstyn. T., and M. D. McCulloch. Scalable multi-agent reinforcement learning for distributed control of residential energy flexibility. *Applied Energy*, 314:118825, 2022.

[214] Jakob N. Foerster, G. Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 2974–2982, 2018.

[215] Ryan Lowe, YI WU, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[216] J. K. Gupta, M. Egorov, and M. Kochenderfer. Cooperative Multi-agent Control Using Deep Reinforcement Learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10642 LNAI:66–83, 2017.

[217] T. Rashid. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning Tabish. *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[218] Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 10199–10210, 2020.

[219] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. QPLEX: Duplex dueling multi-agent Q-learning. In *International Conference on Learning Representations*, 2021.

[220] Jian H., Seth A. H., Haibin W., and Shih-wei L. QR-MIX: distributional value function factorisation for cooperative multi-agent reinforcement learning. *CoRR*, abs/2009.04197, 2020.

[221] W. Qiu. Rmix: Learning risk-sensitive policies for cooperative reinforcement learning agents. 2021.

[222] Bei Peng, Tabish Rashid, Christian Schroeder de Witt, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Boehmer, and Shimon Whiteson. Facmac: Factored multi-agent centralised policy gradients. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12208–12221. Curran Associates, Inc., 2021.

[223] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms in Cooperative Tasks. (NeurIPS), 2020.

[224] P. Sunehag, G. Lever, N. Sonnerat, and M. Jaderberg. Value-Decomposition Networks For Cooperative Multi-Agent Learning. 2012.

[225] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim G.J. Rudner, Chia Man Hung, Philip H.S. Torr, Jakob Foerster, and Shimon Whiteson. The StarCraft multi-agent challenge. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, 4(NeurIPS):2186–2188, 2019.

[226] Rihab Gorsane, Omayma Mahjoub, Ruan de Kock, Roland Dubb, Siddarth Singh, and Arnu Pretorius. Towards a Standardised Performance Evaluation Protocol for Cooperative MARL. (NeurIPS):1–43, 2022.

[227] Jose R Vazquez-Canteli, Sourav Dey, Gregor Henze, and Zoltan Nagy. CityLearn: Standardizing Research in Multi-Agent Reinforcement Learning for Demand Response and Urban Energy Management. (figure 1), 2020.

[228] Shary Heuninckx, Maarja Meitern, Geert te Boveldt, and Thierry Coosemans. Practical problems before privacy concerns: How European energy community initiatives struggle with data collection. *Energy Research and Social Science*, 98(September 2022), 2023.

[229] Department for Transport. National Travel Survey 2002-2020, 2021.

[230] Tanveer Ahmad, Dongdong Zhang, Chao Huang, Hongcai Zhang, Ningyi Dai, Yonghua Song, and Huanxin Chen. Artificial intelligence in sustainable energy industry: Status Quo, challenges and opportunities. *Journal of Cleaner Production*, 289:125834, 2021.

[231] B Nijenhuis, Sjoerd C Doumen, Jens Hönen, and Gerwin Hoogsteen. Using mobility data and agent-based models to generate future e-mobility charging demand patterns. In *CIRED Porto Workshop 2022: E-mobility and power distribution systems*, volume 2022, pages 214–218. IET, 2022.

[232] Eoghan McKenna, Murray Thomson, and John Barton. CREST Demand Model.

[233] Chi Zhang, Sanmukh R. Kuppannagari, Rajgopal Kannan, and Viktor K. Prasanna. Generative Adversarial Network for Synthetic Time Series Data Generation in Smart Grids. *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids, SmartGridComm 2018*, pages 1–6, 2018.

[234] M. Imran Azim, Wayes Tushar, and Tapan K. Saha. Investigating the impact of p2p trading on power losses in grid-connected networks with prosumers. *Applied Energy*, 263:114687, 2020.

[235] C. Coffrin, P. Van Hentenryck, and R. Bent. Approximating line losses and apparent power in AC power flow linearizations. *IEEE Power and Energy Society General Meeting*, pages 1–8, 2012.

[236] R. Dufo-López, J. M Lujano-Rojas, and J. L. Bernal-Agustín. Comparison of different lead–acid battery lifetime prediction models for use in simulation of stand-alone photovoltaic systems. *Applied energy*, 115:242–253, 2014.

[237] ISO. Calculation of Energy Use for Space Heating and Cooling ISO/FDIS 13790:2007(E), 2007.

[238] M. Farivar, C. R. Clarke, S. H. Low, and K. M. Chandy. Inverter var control for distribution systems with renewables. In *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pages 457–462. IEEE, 2011.

[239] M.E. Baran and F.F. Wu. Optimal capacitor placement on radial distribution systems. *IEEE transactions on power delivery*, 4(1):725–734, 1989.

[240] Na Li, Lijun Chen, and Steven H. Low. Exact convex relaxation of OPF for radial networks using branch flow model. *2012 IEEE 3rd International Conference on Smart Grid Communications, SmartGridComm 2012*, pages 7–12, 2012.

[241] Eleni Stai, Lorenzo Reyes-Chamorro, Fabrizio Sossan, Jean Yves Le Boudec, and Mario Paolone. Dispatching stochastic heterogeneous resources accounting for grid and battery losses. *IEEE Transactions on Smart Grid*, 9(6):6522–6539, 2018.

[242] Leon Thurner, Alexander Scheidler, Florian Schafer, Jan Hendrik Menke, Julian Dollichon, Friederike Meier, Steffen Meinecke, and Martin Braun. Pandapower - An Open-Source Python Tool for Convenient Modeling, Analysis, and Optimization of Electric Power Systems. *IEEE Transactions on Power Systems*, 33(6):6510–6521, 2018.

[243] F. Charbonnier, T. Morstyn, and M. D. Mcculloch. Coordination of resources at the edge of the electricity grid : Systematic review and taxonomy. *Applied Energy*, 318(April):119188, 2022.

[244] Tanveer Ahmad, Huanxin Chen, Yabin Guo, and Jiangyu Wang. A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review. *Energy and Buildings*, 165:301–320, 2018.

[245] Andreas I. Elombo, Thomas Morstyn, Dimitra Apostolopoulou, and Malcolm D. McCulloch. Residential load variability and diversity at different sampling time and aggregation scales. *2017 IEEE AFRICON: Science, Technology and Innovation for Africa, AFRICON 2017*, pages 1331–1336, 2017.

[246] Edward O'Dwyer, Indranil Pan, Salvador Acha, and Nilay Shah. Smart energy systems for sustainable smart cities: Current developments, trends and future directions. *Applied Energy*, 237(October 2018):581–597, 2019.

[247] Xiangru Lian, Ce Zhang, Huan Zhang, Cho Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 2017-Decem(1):5331–5341, 2017.

[248] R. Wardle. Dataset (TC1a): Basic Profiling of Domestic Smart Meter Customers, 2014.

[249] R. Wardle. Dataset (TC5): Enhanced Profiling of Domestic Customers with Solar Photovoltaics (PV), 2014.

[250] C. Crozier, D. Apostolopoulou, and M. McCulloch. Numerical analysis of national travel data to assess the impact of UK fleet electrification. *20th Power Systems Computation Conference, PSCC 2018*, pages 1–7, 2018.

[251] Jouni Peppanen, Xiaochen Zhang, Santiago Grijalva, and Matthew J. Reno. Handling bad or missing smart meter data through advanced data imputation. *2016 IEEE Power and Energy Society Innovative Smart Grid Technologies Conference, ISGT 2016*, pages 1–5, 2016.

[252] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, mar 1982.

[253] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. 2014.

[254] Kaichao You, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. How Does Learning Rate Decay Help Modern Neural Networks? 2019.

[255] Johannes Lederer. Activation Functions in Artificial Neural Networks: A Systematic Overview. pages 1–42, 2021.

[256] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[257] Eoin Brophy, Zhengwei Wang, Qi She, and Tomás Ward. Generative Adversarial Networks in Time Series: A Systematic Literature Review. *ACM Computing Surveys*, 55(10), 2023.

[258] Stephanie L. Hyland, Cristóbal Esteban, and Gunnar Rätsch. Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs. 2017.

[259] D. Hirst. Commons Briefing Paper SNO5927: Carbon Price Floor (CPF) and the price support mechanism, 2018.

[260] Global Modeling and Assimilation Office (GMAO). Merra-2 inst1_2d_asm_Nx: 2d,1-hourly, instantaneous, single-level, assimilation, single-level diagnostics v5.12.4, 2015.

[261] Octopus Energy. Octopus Energy API, 2019.

[262] J. Brown, J. Chambers, and A. Rogers. SMITE : Using Smart Meters to Infer the Thermal Efficiency of Residential Homes. In *The 7th ACM International Conference on Systems for Energy- Efficient Buildings, Cities, and Transportation (BuildSys '20)*, 2020.

[263] Nissan Intelligent Mobility. Nissan leaf.

[264] National Grid ESO, Environmental Defense Fund Europe, University of Oxford Department of Computer Science, and WWF. Carbon Intensity API, 2020.

[265] K. P. Schneider, B. A. Mather, B. C. Pal, C. W. Ten, G. J. Shirek, H. Zhu, J. C. Fuller, J. L. R. Pereira, L. F. Ochoa, L. R. de Araujo, R. C. Dugan, S. Matthias, S. Paudyal, T. E. McDermott, and W. Kersting. Analytic considerations and design basis for the IEEE distribution test feeders. *IEEE Transactions on Power Systems*, PP(99):1–1, 2017.

[266] O. Vinyals. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(November), 2019.

[267] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational information maximizing exploration. *Advances in Neural Information Processing Systems*, 0:1117–1125, 2016.

[268] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Comparative Evaluation of Multi-Agent Deep Reinforcement Learning Algorithms. 2020.

[269] Stephanie C. Y. Chan, Samuel Fishman, John Canny, Anoop Korattikara, and Sergio Guadarrama. Measuring the Reliability of Reinforcement Learning Algorithms. pages 1–36, 2020.

[270] P. Virtanen. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[271] T. Rashid, G. Farquhar, B. Peng, and S. Whiteson. Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 2020-December, 2020.

[272] Frans A Oliehoek and Christopher Amato. *A concise introduction to decentralized POMDPs*. Springer, 2016.

[273] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[274] Steven B. Damelin and Willard Miller. *The mathematics of signal processing*. Cambridge texts in applied mathematics ; 48. Cambridge University Press, Cambridge, 2012.

[275] Mikolaj Binkowski, Gautier Marti, and Philippe Donnat. Autoregressive convolutional neural networks for asynchronous time series. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 580–589. PMLR, 10–15 Jul 2018.

[276] Zhicheng Cui, Wenlin Chen, and Yixin Chen. Multi-scale convolutional neural networks for time series classification, 2016.

[277] Barbara Mukami Maweu, Rittika Shamsuddin, Sagnik Dakshit, and Balakrishnan Prabhakaran. Generating healthcare time series data for improving diagnostic accuracy of deep neural networks. *IEEE Transactions on Instrumentation and Measurement*, 70:1–15, 2021.

[278] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013.

[279] Xiang Gao. Deep reinforcement learning for time series: playing idealized trading games, 2018.

[280] T. Hester. Deep q-learning from demonstrations. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 3223–3230, 2018.

[281] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[282] S. J. Wright. *Primal-Dual Interior-Point Methods*. Society for Industrial and Applied Mathematics, 1997.

[283] K. Baker. A learning-boosted quasi-newton method for ac optimal power flow. *arXiv preprint arXiv:2007.06074*, 2020.

[284] Matthew Deakin, Thomas Morstyn, Dimitra Apostolopoulou, and Malcolm D McCulloch. Voltage control loss factors for quantifying dg reactive power control impacts on losses and curtailment. *IET Generation, Transmission & Distribution*, 16(10):2049–2062, 2022.

[285] Flora Charbonnier, Thomas Morstyn, and Malcolm McCulloch. *Active Players in Local Energy Markets*, pages 71–111. Springer International Publishing, Cham, 2023.

[286] Iacopo Savelli and Thomas Morstyn. Better together: Harnessing social relationships in smart energy communities. *Energy research and social science*, 78:102125, 2021.

[287] Jessica M. Nolan, P. Wesley Schultz, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius. Normative social influence is underdetected. *Personality and Social Psychology Bulletin*, 34(7):913–923, 2008.

[288] Council of European Energy Regulators. Regulatory Aspects of Self- Consumption and Energy Communities CEER Report. Technical Report June, 2019.

[289] Official Journal of the European Union. DIRECTIVE (EU) 2019/944 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 5 June 2019 on common rules for the internal market for electricity and amending Directive 2012/27/EU. 2019.

[290] Flora Charbonnier, Thomas Morstyn, and Malcolm McCulloch. Home electricity data generator (hedge): An open-access tool for the generation of electric vehicle, residential demand, and pv generation profiles, 2023.

[291] Shaohui Liu, Chengyang Wu, and Hao Zhu. Topology-aware graph neural networks for learning feasible and adaptive ac-opf solutions. *IEEE Transactions on Power Systems*, pages 1–11, 2022.

[292] Rahul Nellikkath and Spyros Chatzivasileiadis. Physics-informed neural networks for minimising worst-case violations in dc optimal power flow. In *2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 419–424, 2021.

[293] Rahul Nellikkath and Spyros Chatzivasileiadis. Minimizing worst-case violations of neural networks, 2022.

[294] Rahul Nellikkath and Spyros Chatzivasileiadis. Enriching neural network training dataset to improve worst-case performance guarantees, 2023.

[295] Samuel Chevalier and Spyros Chatzivasileiadis. Global performance guarantees for neural network models of ac power flow, 2023.

[296] British Standards. Heating systems in buildings. Method for calculation of the design heat load. *Ics 91.140.10*, (January):1–89, 2009.

[297] V. Becker, W. Kleiminger, V. Coroamă, and F. Mattern. Estimating the savings potential of occupancy-based heating strategies. *Energy Informatics*, 1(S1), 2018.

[298] W. Powell. *Approximate dynamic programming: solving the curses of dimensionality.* Wiley series in probability and statistics. J. Wiley & Sons, Hoboken, N.J., 2nd ed. edition, 2011.

[299] Quasar. Wallbox quasar: The first bidirectional charger for your home.

[300] David Young. Wallbox quasar ev charging in the uk. *The Institute of Automotive Engineer Assessors*, Sep 2020.

[301] HOMER Energy. HOMER Pro 3.14 User Manual, 2020.

[302] W. Schram. Empirical evaluation of V2G round-trip efficiency. *SEST 2020 - 3rd International Conference on Smart Energy Systems and Technologies*, (October), 2020.

[303] R. Tonkoski, D. Turcotte, and T. H.M. El-Fouly. Impact of high PV penetration on voltage profiles in residential neighborhoods. *IEEE Transactions on Sustainable Energy*, 3(3):518–527, 2012.

[304] UK Legislation. Electricity safety, quality and continuity regulations 2022, part VII, Regulation 27, 2002.