Automated Coordination of Distributed Energy Resources using Local Energy Markets and Reinforcement Learning

by

Daniel Christopher May

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Software Engineering and Intelligent Systems

Department of Electrical and Computer Engineering
University of Alberta

Abstract

The conventional unidirectional model of the electricity grid operations is no longer sufficient. The continued proliferation of distributed energy resources and the resultant surge in net load variability at the grid edge necessitates deploying adequate demand response methods.

This thesis proposes, investigates, and demonstrates the Autonomous Local Energy eXchange (ALEX), an indirect demand response mechanism grounded in the principles of transactive energy. ALEX operates as a fully automated, decentralized, and economy-driven local energy market with the overarching objective of reducing net load variability on a community level to enhance grid operability. ALEX strongly distinguishes itself from schedule-based approaches commonly utilized for indirect demand response in how it addresses the challenges of interest alignment and end-user participation. The alignment of end-user and grid stakeholder interests is achieved through the market mechanism, which incentivizes pricing in relation to the current timestep's supply/demand ratio. To facilitate broad end-user participation in the face of such a granular incentive signal, ALEX relies on model-free automation through deep reinforcement learning.

The thesis employs a reductionist approach to navigate the complex dynamics of this interconnected system. It formulates three primary research goals, addressed through corresponding chapters.

Chapter 2 explores the challenges of economy-driven transactive energy, focusing on designing an appropriate local energy market mechanism. Through classification driven experiments, a market mechanism is identified that strongly incentivizes pricing in relation to the supply/demand ratio. This provides an effective solution to the alignment problem between grid stakeholders and electricity end-users.

Chapter 3 develops a benchmarking approach for local energy markets, confirming the central hypothesis of emergent, community-level variability reduction within ALEX. ALEX significantly outperforms baseline approaches, demonstrating its capability to enable community-wide coordination of distributed energy resources. The benchmarking process addresses broader research gaps in the current literature related to Local Energy Markets.

Chapter 4 concludes the thesis by training deep reinforcement learning agents to achieve near-optimal performance on ALEX. An augmented proximal policy optimization algorithm demonstrates the ability to produce a convergent set of policies close to a Nash equilibrium. The resulting policies reduce community-level variability across several timescales without information sharing between agents and without access to future information.

In summary, this thesis advances the state of the art in indirect demand response by introducing and demonstrating ALEX. ALEX's decentralized, autonomous nature positions it as a robust solution to the challenges posed by the growing adoption of distributed energy resources, aligning with the Smart Grid's principles of intelligent asset integration for efficient and reliable grid operations.

Preface

The research detailed in this thesis unfolded under the guidance of Dr. Petr Musilek at the Energy Digitization Lab (ENTAIL) and Dr. Matthew Taylor at the Intelligent Robot Learning Lab (IRL Lab). Chapters 2 to Chapter 4 have been published or are submitted as [1], [2], and [3], respectively:

> [1]: Steven Zhang, Daniel May, Mustafa Gül, Petr Musilek, Reinforcement learning-driven local transactive energy market for distributed energy resources, Energy and AI, Volume 8, 2022, 100150, ISSN 2666-5468, https://doi.org/10.1016/j.egyai.2022.100150.
>
> [2]: Daniel May and Petr Musilek, Transactive Local Energy Markets Enable Community-Level Resource Coordination Using Individual Rewards, 2024, Submitted to IEEE Access, Preprint available at https://arxiv.org/abs/2403.15617
>
> [3]: Daniel May, Matthew Taylor and Petr Musilek, Decentralized Coordination of Distributed Energy Resources through Local Energy Markets and Deep Reinforcement Learning, 2024, Submitted to Energy and AI

As the primary contributor to [1], Dr. Steven Shida Zhang played a pivotal role in shaping the concept of ALEX, managing the codebase, experimental analyses, and the overall crafting of the manuscript. In parallel, my contributions to this publication centered around developing the game-theoretic model for ALEX, implementing the reinforcement learning algorithm, and essential elements within the background, experimental design, and discussion sections of the manuscript.

As the primary author of the studies [2, 3], my responsibilities encompassed conceptualization, implementation, experimental design, analysis, and manuscript composition. Dr. Petr Musilek, in addition to his role as a supervising author, provided valuable mentorship and contributed to finalizing the manuscript drafts. Dr. Matthew Taylor guided the research from its conceptualization for both contributions, serving as a supervising author for [3].

This thesis presents a tightly focused narrative toward a clear research objective and does not encompass my engagement in adjacent research areas. One example of such research is the collaboration with Elizaveta Kharlova:

> E. Kharlova, D. May and P. Musilek, Forecasting Photovoltaic Power Production using a Deep Learning Sequence to Sequence Model with Attention, 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-7, doi: 10.1109/IJCNN48605.2020.9207573.

Elizaveta Kharlova's meticulous analysis of attention mechanisms significantly enhanced the article's discussion and background sections, complementing her thesis. My contributions focused on the original investigation into sequence-to-sequence models for probabilistic forecasting, experiment design, formulation of the evaluation methodology, and development of the theoretical framework underpinning the research.

Acknowledgements

The research presented here was shaped and influenced by countless individuals. First and foremost, however, I would like to express my deepest gratitude to my wonderful wife, Sonia. Your radiant love, patience, spoken and silent encouragement shone a light even during the darkest of times.

Acknowledging the indispensable guidance and mentorship received throughout my academic journey, I express my sincere thanks to Dr. Petr Musilek and Dr. Matthew Taylor. Petr's unwavering patience and stoic support persisted even when faced with my stubbornness, and his continued respect for my autonomy in choosing research topics was pivotal. Matthew's consistent support, humor, openness, reliability, and competence make him the perfect mentor. I am truly grateful to graduate under such guidance.

Special mention goes to Dr. Nathan Deisman, a friend and mentor who continues to be a source of inspiration, guidance, and calming influence beyond the limitations of words, my parents and my brother for their unconditional love and all that we are to each other.

Last but not least, Dr. Steven Zhang and Peter Atrazhev, my true 'partners in crime'. You always have my back. This thesis is built from our late-night discussions, struggles and hopes. I am eager to see where this journey fuelled by our collective madness leads us next - fail fast (sometimes slow) but always fail forward (more often sideways).

Table of Contents

List of Tables

List of Figures

Abbreviations

ALEX Autonomous Local Energy eXchange.

BESS Battery Energy Storage System.

DERMS DER Management System.

DERs Distributed Energy Resources.

DP Dynamic Programming.

DQN Deep Q Networks.

DR Demand Response.

DRL Deep Reinforcement Learning.

EV Electric Vehicle.

LEM Local Energy Market.

LSTM Long Short-Term Memory.

MDP Markov Decision Process.

MPC Model Predictive Control.

PPO Proximal Policy Optimization.

PV Photovoltaics.

RL Reinforcement Learning.

SoC State of Charge.

TE Transactive Energy.

TOU Time Of Use Pricing.

Chapter 1

Introduction

Climate change stands as an unequivocal global challenge, necessitating collaborative, multi-generational endeavors. Addressing this issue mandates a dual focus on sustainable energy consumption and accommodating the escalating energy demands propelled by technological advancements. In the face of this vast spectrum of adaptations, the maintenance of reliable and efficient operations is paramount.

The electricity grid is conventionally treated as a unidirectional, hierarchical graph. A powerplant generates electricity at the graph's origin, from where a cascade of levels with increasing local resolution distributes it to the end-user at the grid edge to satisfy their load demand. These end-users are traditionally regarded as passive consumers with reasonably predictable consumption patterns. Unit-commitment and economic dispatch algorithms schedule power plants while the grid's configuration adjusts to the implied power flow. Mechanisms such as spinning reserves and ancillary services mitigate deviations from forecasts, ensuring equilibrium between energy sources and sinks. The directed graph structure of this model focuses efforts toward higher voltage levels and major nodes, optimizing adjustments in an effective, centralized manner.

The energy sustainability efforts materialize through many avenues. Electrification in the form of electric vehicles and electric heating shifts the primary mode of energy consumption. Concurrently, consumer end-users evolve into prosumers propelled by the economic viability of residential electricity generation in photovoltaic panels, wind turbines, and geothermal energy [4, 5]. These adaptations drastically alter end-user behavior and occur at a pace clearly surpassing the typical infrastructure upgrade cycles of the electricity grid. Consequently, this challenges the established centralized operational paradigm of the grid. In response, the Smart Grid emerges as a comprehensive 'catch-all' solution concept, characterized through a decentralized autonomous operation that intelligently integrates and leverages all available system participants and assets to ensure an efficient and reliable grid [6].

1.1 Distributed Energy Resources

One of the foremost ambitions of the Smart Grid is addressing the pronounced surge in the variability of the end-user netload, which encompasses the composite effects of intermittency and other net load volatilities. In this context, intermittency is strictly defined as the fluctuations in net load attributed to uncontrollable factors. While photovoltaic and wind energy are inherently intermittent, electrified assets such as HVAC systems and electric vehicles also induce instantaneous load demand surges. The result is a marked amplification of day-to-day and moment-to-moment variability of end-user energy usage. This intensifies the steepness and unpredictability of net load ramps, turning the maintenance of electricity grid operations increasingly intricate and brittle. The shape of the 'duck-curve', depicted in Figure 1.1 effectively encapsulates this significantly increased variability.

The adoption of the aforementioned assets unfolds disparately across the electricity grid, with certain areas advancing more rapidly based on individual values and financial capacities. The uneven distribution of these distributed energy resources (DERs) compounds operational challenges, particularly in the absence of comprehensive monitoring infrastructure at the grid edge.

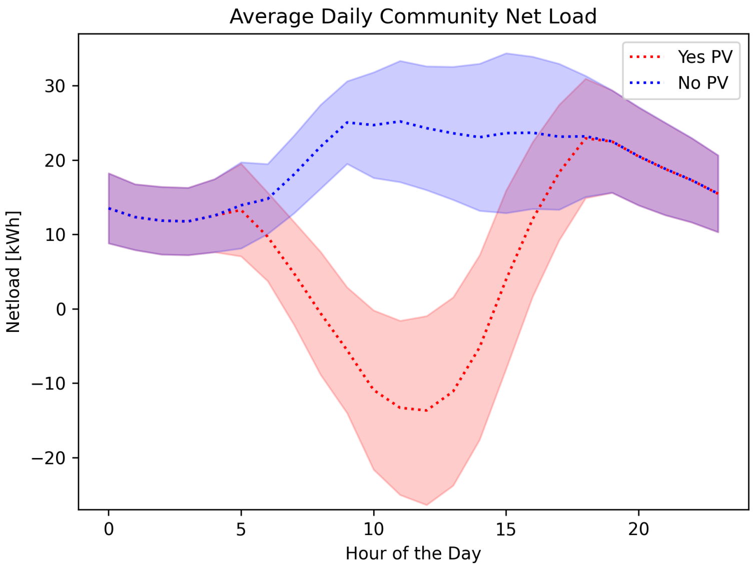

Figure 1.1: Average day net-load profiles for a community with and without residential photovoltaics (PV). Shaded areas depict variance bands. With PV, the community net load curve exhibits a clear 'duck-curve' pattern and elevated variability.

1.2 Demand Response

The first-principles approach to address these challenges involves transforming end-users and their DERs into active contributors to grid stabilization and operation. While the term "Demand Response" (DR) lacks a universally adopted, strict definition, it generally refers to the set of mechanisms employed to shape end-user load demand via signaling to maintain or enhance grid operations [7, 8]. DR approaches can be broadly classified into two categories: direct DR and indirect DR.

Direct DR involves the utilization of a control signal, allowing grid stakeholders access to energy assets to shape load demand. Albeit effective, direct DR faces challenges stemming from inherent conflicts of interest between asset owners and grid operators. Additionally, its tendency towards centralized approaches and the requisite communication and data exchange infrastructure may prove inadequate for addressing the decentralized nature of DERs and their diverse adoption rates [9].

In contrast, indirect DR relies on incentive signals, often in monetary form, to guide end-users' behavior by aligning their interests with grid stakeholder objectives. The effectiveness of an incentive signal hinges on its value perceived by end-users and its ability to elicit the desired behavioral response from the perspective of grid stakeholders. End-users are typically motivated by the goal of minimizing electricity costs and enhancing their well-being, while grid stakeholders prioritize the stable operation of the electricity grid and efficient revenue generation. Once communicated, garnering a substantial end-user response toward an incentive signal becomes critical to achieve the intended effects. This poses a challenge for non-automated, indirect DR programs [8, 9].

1.3 Transactive Energy

Indirect DR signals vary over a spectrum from fixed schedules like Time-of-Use to real-time pricing derived from the instantaneous state of the electricity grid at the moment of transaction [9]. Traditionally, this incentive signal is formulated through model predictive control (MPC), relying on behavioral models and a centralized information processing structure, and then communicated as a price schedule. For sustained effectiveness, DR must adapt to current trends, evolve to accommodate a myriad of participants, and account for the distributed nature of DERs.

This imperative motivates the emergence of the transactive energy (TE) framework. The GridWise Architecture Council defines TE as "the use of a combination of economic and control techniques to improve grid reliability and efficiency" [10]. TE introduces a market-oriented incentive delivery mechanism that fundamentally distinguishes it from conventional MPC-derived pricing schedules. In contrast to the centralized nature of the electricity grid's wholesale market, TE emphasizes decentralization and automation of transactions and control, aligning elegantly with the notion of the Smart Grid [9].

Various approaches exist to assign value to transactions within the TE framework. Some studies adopt power flow-driven pricing based on one or more power system performance metrics [11–14]. While this ensures the effectiveness of the incentive signal concerning the selected metrics, grid stability is a multi-objective concept with variability occurring on various time scales. The computational expense of calculating an exhaustive set of informing metrics reverts to an MPC-driven incentive scheme.

Other studies employ economy-driven TE, which derives price from economic and market-centric considerations [15–18]. An economy-driven TE system offers the advantage of not relying on additional monitoring infrastructure, being computationally less expensive, and aligning better with the original intentions behind the TE proposal. However, the main challenge lies in designing the market appropriately to incentivize behavior that reduces variability across all time scales.

In recent literature, the local energy market (LEM) has emerged as a promising concept for the implementation of TE at the grid edge [19, 20].

1.4 Automation of Demand Response

While one might expect heightened end-user engagement with a less variable scheduling approach like Time-of-Use, pilot studies indicate that the response to any incentive signal remains notably low in the absence of facilitating factors such as automation [8, 9, 21].

To address this participation challenge, the majority of DR mechanisms proposed in the literature rely on the MPC framework: leveraging a behavioral model to generate a forecast, which then informs a search algorithm that devises a target schedule. The primary advantages of MPC include a degree of explainability and a soft guarantee of optimality. However, drawbacks include reliance on experts, extended execution times and brittleness due to the layered model and forecast structure. The computational demands of the MPC loop impose a lower limit on the step size, and the framework favors centralized approaches. Moreover, compared to model-free approaches, MPC systems struggle to adapt to changes in the controlled system [22].

Developments in machine learning, deep learning, and specifically, deep reinforcement learning (DRL) open up such model-free control approaches as a viable alternative to MPC [23]. DRL has demonstrated effectiveness in complex system control tasks, emerging as the state-of-the-art for robotics [24] and general process control applications [25]. Implementing control through DRL offers a pathway to decentralized participation automation, where each end-user could be automated through their individual DRL agent. Such advantages are further reinforced by general considerations on performance scaling, such as outlined in Rich Sutton's Bitter Lesson [26]. Given that the advantages of the DRL framework align with the Smart Grid vision and correspond to some of the challenges faced by DR in the context of DERs, an expanding body of literature is exploring its applications [27–29].

1.5 Research Objectives and Outline

This thesis endeavors to formulate an indirect DR system aligned with the Smart Grid paradigm and the previously outlined developments. To this end, the research presented within this thesis proposes, investigates, and demonstrates the Autonomous Local Energy eXchange (ALEX): a fully decentralized, economy-driven, TE-based LEM in which end-user participation and DER control are automated via DRL agents. Aligning end-user and grid stakeholder interests through a LEM results in a far more granular incentive signal than one derived from a schedule-based mechanism. Sufficient end-user participation is consequently ensured using decentralized, model-free control through DRL agents.

The central challenge in achieving this thesis goal arises from the intricate dynamics of the interconnected system interplay between the LEM and its automated participants. Simultaneously developing both components poses significant challenges, hindering the design of robust experiments and increasing the likelihood of deviant, high-performant solutions, as observed in experiments like OpenAI's Hide and Seek [30]. Thus, a reductionist approach is adopted, aiming to minimize the interplay between system sub-components, providing clearer insight and informing more precise experimental design. To demonstrate ALEX, the following research objectives must be met:

- Design of a suitable LEM
- Development of an adequate DRL algorithm and training routine
- Establishment of an appropriate benchmarking and evaluation process

As highlighted by Mengelkamp et al. [20], decoupling the effects of the LEM from the behavior of the automating agent is an extremely intricate task. The terms 'suitable', 'adequate', and 'appropriate' in the objective formulation denote achieving a specific objective without hindering the remaining ones while enabling ALEX to exhibit the desired performance attributes as a DR system. Drawn from corresponding academic contributions published as journal articles, Chapters 2 to 4 each address one of these research objectives. This process enables clearer insights into the interactions within the system, setting the stage for a comprehensive understanding of ALEX.

Chapter 2 is published as "Reinforcement learning-driven local transactive energy market for distributed energy resources" in Energy and AI [1]. This contribution forms the initial proposal of ALEX. It explores the challenges of economy-driven TE, focusing on designing a LEM settlement mechanism that efficiently solves the alignment problem between grid stakeholders and electricity end-users. It identifies a set of properties necessary for LEM to achieve this efficiently in the presence of RL agents. Criteria for classifying a double auction market are established, and a set of corresponding LEMs are constructed and tested with RL agents in a bandit setting. This effort identifies a market mechanism for ALEX that strongly incentivizes pricing in relation to the settlement time step's supply and demand ratio. This informs the hypothesis that such a market mechanism could be sufficient to encourage behaviors that emergently reduce community-level variability.

In Chapter 3, submitted to IEEE Access as "Transactive Local Energy Markets Enable Community-Level Resource Coordination Using Individual Rewards", an evaluation procedure is developed to assess ALEX's impact on variability across various time horizons in the presence of residential load shifting capabilities. Formulating ALEX as a discrete markov decision process (MDP), this approach employs iterative best response and dynamic programming to derive a strong, near-optimal performance baseline. This contribution validates the hypothesis of emergent, community-level variability reduction through ALEX and serves as an evaluation benchmark for any trained DRL agents.

Chapter 4, submitted to Energy and AI as "Decentralized Coordination of Distributed Energy Resources through Local Energy Markets and Deep Reinforcement Learning" [3], concludes the research. The main contribution is the development of a DRL algorithm producing a converging set of agents within ALEX, demonstrating the desired emergent DR properties. Built on the popular proximal policy optimization algorithm [31], the approach implements several general algorithm improvements [32–34] Notably, the achieved performance level, close to the near-optimal baseline established in Chapter 3, is accomplished without access to future information, affirming ALEX's capabilities as originally intended.

In its entirety, this thesis documents contributions to the development and demonstration of ALEX, a fully data-driven, model-free DR system significantly reducing net-load variability on a community level. Crucially, ALEX operates in a decentralized, autonomous manner with only building-level information, offering the necessary scalability and adaptability for the future electricity grid.

Chapter 2

Reinforcement Learning-Driven Local Transactive Energy Market for Distributed Energy Resources

2.1 Introduction

Demand response (DR) techniques have become popular means to increase the value of distributed energy resources (DER), such as rooftop solar, while mitigating the negative effects of their intermittent nature. DR methods can be direct or indirect. Indirect DR aims to change customer behavior using an incentive signal, usually through monetary means [35–37]. Direct DR grants grid operators immediate control to perform grid balancing. As DER adoption continues, centralized approaches to DR will encounter scalability barriers [9, 35, 38, 39], and more granular and robust control will be necessary due to the increased intermittency and stochasticity of supply and demand. Despite the effort to tackle some of these challenges [40–42], it is increasingly clear that alternative, decentralized solutions, must be explored [43]. In this context, transactive energy (TE) is gaining popularity as a design framework for decentralized DR [9, 39]. The U.S. Department of Energy Gridwise Architecture Council defined TE as "a system of economic and control mechanisms that allows the dynamic balance of supply and demand across the entire electrical infrastructure using value as a key operational parameter" [44]. A locally constrained TE system is often referred to as local energy market (LEM). In lieu of a formal definition of LEM [20, 45], we adopt the description developed by Mengelkamp et al. [20], i.e. “a market platform for trading locally generated (renewable) energy among residential customers within a geographically and socially close community. Supply security is ensured through connections to a superimposed energy system.”

High DR participation rates, especially within a LEM, can only be maintained with automation. Expert-designed, rule based systems have initially been considered for this purpose. However, learning-based approaches are now preferred, mainly because of their robustness and scalability. However, the vast majority of existing approaches that apply learning methods to LEM do not tailor the market mechanism to the algorithm used for automation. Given the fact that most established LEM mechanisms were designed for human participants or rule-based system automation [46], this is especially problematic. The result is suboptimal DER utilization, as the LEM is not appropriately adjusted to best leverage the potential of the automation approach used. This is exacerbated for reinforcement learning (RL) agents that can quickly learn to exploit loopholes in competitive-collaborative multi-agent settings [30]. Mengelkamp et al. clearly identify this research gap in their review, stating that “a comprehensive comparison of the impacts of different trading designs (especially market mechanisms) should be carried out. Specifically, the impact of different allocation mechanisms on the market objectives and agent behavior need to be evaluated” [20]. Their later work [47] follows the same reasoning, noting that both agent design and market design influence the resulting system behavior. We argue that the LEM mechanism should be tailored to adequately empower the strengths of their automation methods and to mitigate their potential weaknesses.

To the best of our knowledge, there has been no contribution that explicitly designs experiments to identify the requirements that a specific LEM market mechanism must fulfill to be well compatible with independently-learning RL actors. This article aims to provide a starting point to fill this research gap. We narrowly focus on the following two questions:

- What are the required properties of the LEM settlement mechanism suitable for deployment of RL-based automation?
- Does the resulting market behavior effectively support DR for LEM with high penetration of DER?

To answer these questions, three different settlement mechanisms are examined that cover a set of established criteria for auction environments. The most suitable market design is found by analyzing the agent policies developed for each mechanism. This is followed by modeling the resulting LEM transactions as a dynamic price signal and comparing its economic performance with existing pricing methods.

This article is organized in five sections. Section 2.2 provides the necessary background and describes the related work. Section 2.3 introduces the proposed autonomous local energy exchange (ALEX) and describes it as a stochastic game. Section 2.4 describes two sets of experiments. The first set is designed to identify settlement mechanism suitable for market automation using learning agents. The second set performs an economic analysis of the selected mechanism and compares its performance with several benchmarks. Major conclusions are summarized in Section 2.5, along with possible directions for future work. Appendices present a brief introduction to RL [48] (A.1), overview the principles of net billing (A.2), describe the transactive energy simulator T-REX (A.3), and provide the details of the specific market design used in this article (A.4).

2.2 Background and Related Work

This section provides an in-depth review of the related work, focusing on articles that combine RL with LEM. For a broader context of LEMs, the reader may refer to a general review by Mengelkamp et al. [20], game-theory focused review by Pilz et al. [45], and review of LEM settlement and market mechanisms by Khorasany et al. [46].

2.2.1 Reinforcement Learning for Local Energy Markets

Several authors investigate the combination of RL and dynamic pricing for centralized control. Notably, Kim et al. [36], and Lu et al. [37] develop RL-based approaches for dynamic pricing from the perspective of a service retailer. Both articles address difficulties of predicting participant response to a pricing schedule by mitigating the reliance on accurate customer side information. A Markov decision process is formulated based on customer behaviour models and preferences. A Q-learning agent is trained to simultaneously minimize customer costs and maximize the service provider benefit. The two approaches differ in the formulation of the reward function, which is a major influencing factor in RL algorithms. Lu et al. [37] use a weighted sum of retailer and customers, while Kim et al. [36] use a modelled utility function. Although both proposed approaches successfully implement dynamic pricing strategies without scheduling, they still rely on modeling consumer behavior and preferences via utility functions. Liu et al. [43] address some of these weaknesses by applying deep reinforcement learning (DRL) in a consumer-centric, resource-sharing economy model.

Zhang et al. [49] train an RL agent to manage a community-shared battery and trade its resources on a TE market to maximize economy. The reward function is the economic performance of the battery. The authors show that positive economic benefits can be achieved, even when considering the running costs of the battery.

Foruzan et al. [50] investigate the behavior of independent Q-learning agents, exchanging energy within a microgrid through a LEM. Each agent's goal is to maximize its own profit. Managed DERs include battery energy storage systems, rooftop solar, wind and diesel generators. The participants' stochastic behavior is approximated using random models. The authors investigate several micro grid configurations and perform an in-depth hyperparameter study of the RL algorithm with respect to return, self-sufficiency and fairness.

Zhou et al. [51] combine a fuzzy rule-based system with Q-learning to train agents to exchange energy resources over a peer-to-peer LEM setup whose pricing is directly tied to the ratio of supply and demand. The authors investigate the performance of several community configurations with ranging number of battery energy storage systems and renewable generation assets. They show that such a system setup generally achieves lower bills than TOU and net-billing baselines.

Chen et al. [52] employ a deep Q-network (DQN) variant to automate the interactions of prosumers equipped with battery energy storage system in a LEM. The RL agents' action space consists of four distinct, discrete actions covering buy/sell and charge/discharge operations. The learned policy surpasses an intuitive, rule-based strategy. It also outperforms a pure random policy equivalent to a zero-intelligence agent, originally proposed by Ghode et al. [53] as a baseline for agent competence in automated markets. In another article, Chen et al. [54] investigate the function of Q-learning based energy brokers as LEM consensus mechanism for settlements with profit used as the agent's reward. Using several ablation and sensitivity studies, the authors show that the brokers efficiently learn how to maximize their own profit and the efficiency of the market. A recent article by Jogunola et al. [55] augments the DQN agent using prioritized experience replay [56] to maximize the economic benefits.

Bose et al. [15] focus on emerging participant interaction within a fixed LEM setup under differing levels of DER penetration. The authors demonstrate that RL-based agents in such environment can cause partial energy self-sufficiency to emerge. They also show that the degree of self-sufficiency and the complexity of agent interactions depends on the level of DER penetration within the market. Mengelkamp et al. [16] study three different extensions of the Erev-Roth RL algorithm applied to automate LEM participation. They find that the extensions further increase the self-sufficiency of the LEM when compared to the original Erev-Roth algorithm [57].

Mengelkamp et al. [47] compare a peer-to-peer LEM against a closed book, double-auction LEM with settlement rounds. The authors compare the performance of zero-intelligence agents and "intelligent" agents adopted from Nicolaisen et al. [58] on both LEM designs. They show that all market scenarios offer similar economic advantages, with the peer-to-peer LEM used by intelligent agents slightly outperforming the other variants. However, they also note that using one strategy on different markets results in different price trends, The authors eventually conclude that agent strategy and market design need to be co-developed to guarantee the system's performance.

Harrold et al. [59] use rainbow DRL to learn arbitrage in a microgrid. Lee et al. [60] apply dynamic pricing and DRL to maximize the profits of multiple electric vehicle charging stations. Although not directly related to the approach presented in this article, these studies are excellent references for future research on the use of energy storage within the proposed LEM.

2.3 ALEX: Autonomous Local Energy Exchange

The need to model customer behaviour via utility functions and heavy reliance on forecasts may hinder the robustness and scalability of traditional DR techniques. Furthermore, due to the amount of DERs that are expected to be on-line in the near future, certain infrastructure requirements may pose additional barriers for the effective deployment of DR systems. For example, the communication and computational costs for centralized control may grow too expensive, especially for the high temporal resolutions necessary to fully capture the intermittent behaviour of rooftop solar panels and stochastic use patterns of electric vehicles.

Keeping these challenges in mind, this article proposes a distributed, multi-agent approach combined with a double auction market mechanism. When designing this approach, dubbed ALEX (autonomous local energy exchange), the following assumptions have been made:

- Participants are self-interested and, therefore, prioritize their own economic well-being in the decision making process.
- Participants are willing to defer some decision making regarding interactions with indirect DR measures to automation (e.g., using RL agents).
- Each participating unit is equipped with a smart meter, and a sufficient amount of high-resolution historical data is available to train the RL agents.
- The large-scale electricity grid that customers are connected to is an infinite bus.

2.3.1 Core Concept

Conceptually, ALEX is a behind-the-meter DR technique for a localized community using a double auction market as a coordination mechanism. Market participants are the customers who live within the community. However, this could be expanded to include entities that are only temporarily present, such as electric vehicles. The market employs double auctions with a fixed settlement frequency Δt. For each interval $[t, t + \Delta t)$, participants can communicate their intention to trade energy by submitting bids

$$\text{bid}_t = (q_t^{\text{bid}}, p_t^{\text{bid}}), q \in [0, ..., q_{\max}^{\text{ask}}], p \in [p_{\min}, ..., p_{\max}], \tag{2.1}$$

and asks

$$\text{ask}_t = (q_t^{\text{ask}}, p_t^{\text{ask}}), q \in [0, ..., q_{\max}^{\text{ask}}], p \in [p_{\min}, ..., p_{\max}], \tag{2.2}$$

where bid_t and ask_t communicate the intention to buy or sell energy, respectively. They are represented by tuples consisting of the desired quantity q and desired price p of energy to be exchanged. Quantities are expressed in watt hours (Wh), and can range from 0 to a designated maximum. $q_{\max}^{\text{ask}}$ should be set to accommodate expected maximum generation derived from historical data. Likewise, $q_{\max}^{\text{bid}}$ should be set to accommodate the expected maximum load demand. Similarly, prices in \$ are within a designated window between p_{min} and p_{max}. Bids and asks are settled pairwise at the end of each settlement round, returning a settlement signal m_t to each participant

$$m_t = (q_t^{\text{settlement}}, p_t^{\text{settlement}}). \tag{2.3}$$

More details on the market implementation are provided in A.4. It is important to stress that in this setting, participants both determine the price signals and make energy management decisions through market interactions, whereas most other DR approaches simply have agents react to external price signals.

The settlement signal m_t is represented as a list of tuples containing the settled quantities, and the respective prices. It is important to note that participants only receive information about their settlements and, therefore, do not have access to information on the behaviour of other participants.

After the internal trades are concluded, any excess generation/demand within the community is exchanged with the electricity grid at retail prices. In this article, we assume net billing, a commonly used practice where excess energy is sold to the grid at price $p_{\text{sell}}^{\text{grid}}$ and deficient energy is purchased from the grid for price $p_{\text{buy}}^{\text{grid}}$ that includes fees. The behind-the-meter setup grants the community a window of profitability by deferring fees. This naturally bounds the range of internal market prices as follows

$$p_{\text{min}} = p_{\text{sell}}^{\text{grid}} <= p_{\text{market}} <= p_{\text{buy}}^{\text{grid}} = p_{\text{max}}. \tag{2.4}$$

We hypothesize that, if the interactions between participants are dominated by the law of supply and demand, then ALEX functions as a decentralized, indirect DR tool. The resulting pricing naturally provides economic incentives for all market participants to balance supply and demand. We demonstrate this using the experiments described in Section 2.4.1.

2.3.2 ALEX as a Stochastic Game

To analyze the properties of the proposed approach, the auction and strategic bidding by the actors can be described using a suitable mathematical model. A game-theoretic representation of ALEX can be derived by modelling the interactions of participants as a discounted stochastic game

$$\Gamma \coloneqq (n, L, S, A, P, R) \;\; \forall t \in [0...T], \lambda \in (0...1), \tag{2.5}$$

where n is the number of players, L is the list players of length $|L| = n$, S is the state space, A is the action space, P represents the state transition probabilities, R is the reward function, t is the current time step over the modelling period $[0...T]$, and λ is the discount factor.

Both S and A can be decomposed into n individual components S^i and A^i, as shown below

$$S = S^1 \times ... \times S^n, \tag{2.6}$$

$$A = A^1 \times ... \times A^n. \tag{2.7}$$

Superscript i refers to a specific individual L^i, while the subscript is reserved for time t. Note that the action space A is separated in notation from a specific set of actions a_t at time step t, as in the commonly used RL nomenclature introduced in Section 2.2.

State transition probabilities are defined for any set of actions a_t taken at time step t, as follows

$$\forall a_t : \;\; P(S_{t+1}|S_t, a_t) \coloneqq S_t \rightarrow S_{t+1} \tag{2.8}$$

Analogous to the RL setting, the reward or payoff in the stochastic game at time step t is defined by

$$R_t \coloneqq S \times A \rightarrow r, \tag{2.9}$$

which maps from (S_t, a_t) to a real number $r \in \mathbb{R}$. Similarly, each agent aims to maximize their own return G_t (A.1). Thus, all participants use their individually developed policy π^i (A.4 and A.5), to determine action set a_t^i based on observations from S_t^i.

At each time step, all agents can interact with the market by submitting bids (2.1) and asks (2.2). This leads to the following definition of action

$$a_t^i = (\mathrm{bid}_t^i, \mathrm{ask}_t^i, e_t^i), \tag{2.10}$$

where the additional parameter, e_t^i, is reserved for future expansion of the model, e.g., to define nonmarket actions, such as battery management or thermal load control.

Finally, the state observations for each agent are defined as follows

$$S^i = (d_t^i, g_t^i, m_{t-1}^i), \tag{2.11}$$

where d_t^i and g_t^i are, respectively, the load demand and generation at time t, and m_{t-1}^i are settlements received at time $t-1$.

Note that the transition probabilities P result from the collective actions of all agents. However, due to the pairwise settlement mechanism, market design, and the observation space, P is not fully accessible to L^i. This ensures that the developed model is a truly stochastic game.

At least one stable Nash equilibrium is guaranteed to exist within Γ, as long as n, S and A are finite. This condition can be guaranteed by limiting prices p to a reasonable decimal place accuracy (e.g., 4 or 5 significant digits commonly used in banking). A is logically bounded by the condition previously defined by (2.4). As a result, S must also be finite, and therefore each implementation of ALEX is guaranteed to exhibit at least one stable Nash equilibrium.

2.3.3 Automation using Reinforcement Learning

Since the interaction through the developed stochastic game Γ requires strategic competence, automating the interactions of participants (typically prosumers, but theoretically any grid-connected entity) with the LEM is a reasonable response to the difficulties of accurately modeling customer behaviour. The proposed approach centers around training RL agents to perform market interaction and energy management actions, compensating for nonoptimal human behaviour. RL is theoretically very suitable for this task, as the stochastic game described in the previous subsection is equivalent to a Markov decision process under an established set of criteria, outlined in [61]. The framework developed in this section is set up to be algorithm agnostic. However, the subsequent experiments described in Section 2.4.1 employ independent Q-learning.

In this Article, the reward function, R^i, is formulated as follows

$$R_t^i = (\text{profit}_t^{i,\text{LEM}} + \text{profit}_t^{i,\text{grid}}) - (\text{cost}_t^{i,\text{LEM}} + \text{cost}_t^{i,\text{grid}}), \tag{2.12}$$

where,

$$\text{cost}_t^{i,\text{LEM}} = q_t^{i,\text{settled}-\text{bids}} \times p_t^{i,\text{settled}-\text{bids}}, \tag{2.13}$$

$$\text{profit}_t^{i,\text{LEM}} = q_t^{i,\text{settled}-\text{asks}} \times p_t^{i,\text{settled}-\text{asks}}, \tag{2.14}$$

$$\text{cost}_t^{i,\text{grid}} = q_t^{i,\text{grid}-\text{buy}} \times p_t^{i,\text{grid}-\text{buy}}, \tag{2.15}$$

$$\text{profit}_t^{i,\text{grid}} = q_t^{i,\text{grid}-\text{sell}} \times p_t^{i,\text{settled}-\text{sell}}, \tag{2.16}$$

and,

$$q_t^{i,\text{settled}-\text{bids}} + q_t^{i,\text{grid}-\text{buy}} = d_t^i \text{ (load demand)}, \tag{2.17}$$

$$q_t^{i,\text{settled}-\text{asks}} + q_t^{i,\text{grid}-\text{sell}} = g_t^i \text{ (generation)}. \tag{2.18}$$

Other considerations, such as social welfare costs, are explicitly excluded as the current aim is to study participant and system behaviour using pure economic performance. Nevertheless, these factors may be included in future studies. Since RL agents are trained using high-frequency smart meter data, explicit customer behaviour models can be omitted. This is because a sufficient amount of data can better capture nuanced and individualized customer behavioural patterns, while maintaining scalability.

2.4 Experiments and Discussion

2.4.1 Suitability of Settlement Mechanism

A typical market participant can be represented by a prosumer's home. The home may contain any combination of generation, storage, and controllable loads. For the experiments, the generation and load sources are taken directly from smart meter data. As the focus of this article is to study the LEM's behavior under differing ratios of available supply and demand, profile shaping via load shifting or battery storage are not considered.

Experimental Design

This experiment focuses on the first research question: What are the required properties of the LEM settlement mechanism suitable for the deployment of RL-based automation? The results of this experiment will also show how different market properties influence the policies learned by the RL agents. Since the environment (i.e., the market mechanism and the participation strategies of other agents) plays just as important role as the learning algorithm, it is imperative to establish the most suitable market mechanism for subsequent research and implementation of more complex agent designs, and action and strategy spaces. The scope of the experiment is carefully managed to magnify the influence of the settlement mechanism on the resulting agent policies while providing strong convergence bounds despite ALEX's properties as a partially observable, nonstationary environment. As a reminder, the market design used in this study and the rules of interaction are described in detail in A.4. In this experiment, three different settlement mechanisms with varying market properties are tested:

1. Average-Price (M1): Trades are settled if the bid price is greater than or equal to the ask price. The settlement price is the average of the bid and ask prices.

2. Exact-Match (M2): Sellers and buyers choose bid and ask prices from a list of

available prices. Trades are settled if the bid price equals the ask price.

3. Exact-Price (M3): Trades are settled if the bid price is greater than or equal to the ask price. The buyer buys from the auctioneer at the bid price, and the seller sells to the auctioneer at the ask price.

Any double auction mechanism can be described by the following properties [62]: individual rationality[1], economic efficiency[2], budget balancing[3], and truthfulness[4]. An ideal mechanism satisfies all four properties, but it cannot be realized in practice [62]. Since the design of ALEX and the use of RL agents ensures economic efficiency and individual rationality, the three settlement mechanisms can be differentiated by truthfulness and budget balancing alone, as shown in Table 2.1.

	Market Property Settings			
Mechanism	**Individual rationality**	**Economic efficiency**	**Budget balancing**	**Truthfulness**
M1	Yes	Yes	Strong	False
M2	Yes	Yes	Strong	True
M3	Yes	Yes	Weak	True

Table 2.1: Settlement mechanism properties

Three scenarios with different community supply/demand ratios are evaluated for all considered settlement mechanisms: over-supply (10:1), over-demand (1:10), and perfect balance (i.e. equal supply and demand). Each mechanism is evaluated based on the policies developed by the agents and the resulting market behaviour based

[1] Individual rationality states that no participant should lose money from joining the auction

[2] In an economically efficient system, at the end of all trading, the items should be in the hands of participants who bid the highest value.

[3] There are two variants of budget balancing: weak and strong. In a weak budget balancing system, a portion of the money transferred also goes to the auctioneer; this is in addition to money transfers between participants which are the only type of exchange in a system with strong budget balancing.

[4] The dominant strategy in a truthful market is for the participants to report prices at what they believe should be the true value of the item to be exchanged.

on emerging equilibrium bid, ask, and settlement prices, given the same training curriculum. The goal is to find a market mechanism that follows the law of supply and demand, and is compatible with RL agent learning behavior. Such a mechanism is expected to produce the following results:

- Excess supply case: The generators compete for demand, driving ask prices low with bid prices following.
- Excess demand case: The consumers compete for supply, driving bid prices high with ask prices following.
- Equal supply and demand case: The bid and ask prices converge around the middle of the available price range.
- For all cases: The mean bid, ask, and settlement prices should have low spread.

A set of $n = 4$ learning participants is considered. Two participants with $d^i > g^i$ act as buyers, and the remaining two participants with $d^i < g^i$ act as sellers. Two of each type of participant maintain competition on both sides of the market and should prevent monopolistic behaviour. Steady-state (flat, time-invariant) energy profiles are employed for each agent, with the collective load demand and supply corresponding to the previously given ratios

$$\frac{g^{\mathrm{LEM}}}{d^{\mathrm{LEM}}} = \frac{\sum_i g^i}{\sum_i d^i}. \tag{2.19}$$

The use of steady-state load profiles reduces the agents' task to only finding the equilibrium pricing. This setup collapses the observation space of each agent to a single point and fixes q^i_{bid} or q^i_{ask} to the residual load. This allows to further improve the purity of the experiment by learning only the price policy. From the view of a single agent, this transforms the experiment into a partially observable, nonstationary multi-armed bandit, where the number of arms corresponds to the number of discrete price actions $|p|$. For each individual participant, an independent tabular Q-learning algorithm can be used, with ϵ-greedy exploration policy and learning rate α, as described in A.1. This maintains loose convergence guarantees despite the properties of the resulting environment [63]. For this experiment, α is set to 0.1, γ to 0.98, and exploration rate ϵ to 0.1. Values of ϵ and α are annealed starting from episode 100 to balance exploration and convergence speed, with a multiplier of 0.98 per episode. Under this simple setup, if the agents fail to develop policies that reflect the previously mentioned criteria, the respective mechanisms will be considered infeasible for subsequent use.

This experiment was performed using the T-REX simulator, which is described in A.3. The simulations were run on a workstation with Ryzen 9 3900X processor and 32GB of 3200MHz DDR4 memory. In this specific setup, each episode took approximately 5 minutes. Detailed experimental configurations can be found on the GitHub repository of the project [64].

Results and Discussion

Recall that, in ALEX, participants both determine the price signals and make energy management decisions through market interactions. Therefore, it is important to study the agent actions as well as the resulting settlement prices. Figure 2.1 shows, for settlement mechanism M1, the policies learned for bid and ask prices, as well as the resulting settlement prices as density plots. The bid, ask, and settlement prices are, in general, closely clustered on the expected side of the price range for both unbalanced cases. However, the balanced case reveals a critical problem. While the settlement prices are concentrated in the middle of the price range (as expected), both bid and ask prices diverge to near the extremities. This phenomenon results from the lack of truthfulness of M1. Agents have no incentive to submit the bid and ask prices that correspond to what they believe should be the value of energy (near the settlement price). Since M1 calculates the settlement price as the average of each pair of bid/ask, this strategy increases the chance of reaching a favorable settlement. However, at the same time, it also increases the reward if an opponent follows a truthful strategy. This behavior is evidently the optimal strategy to employ in this scenario. The price divergence is problematic, especially when continuous, unbounded action spaces were used for price selection: exceedingly large bid/ask prices may cause settlement prices to become unstable. Because of this risk, M1 is disqualified.

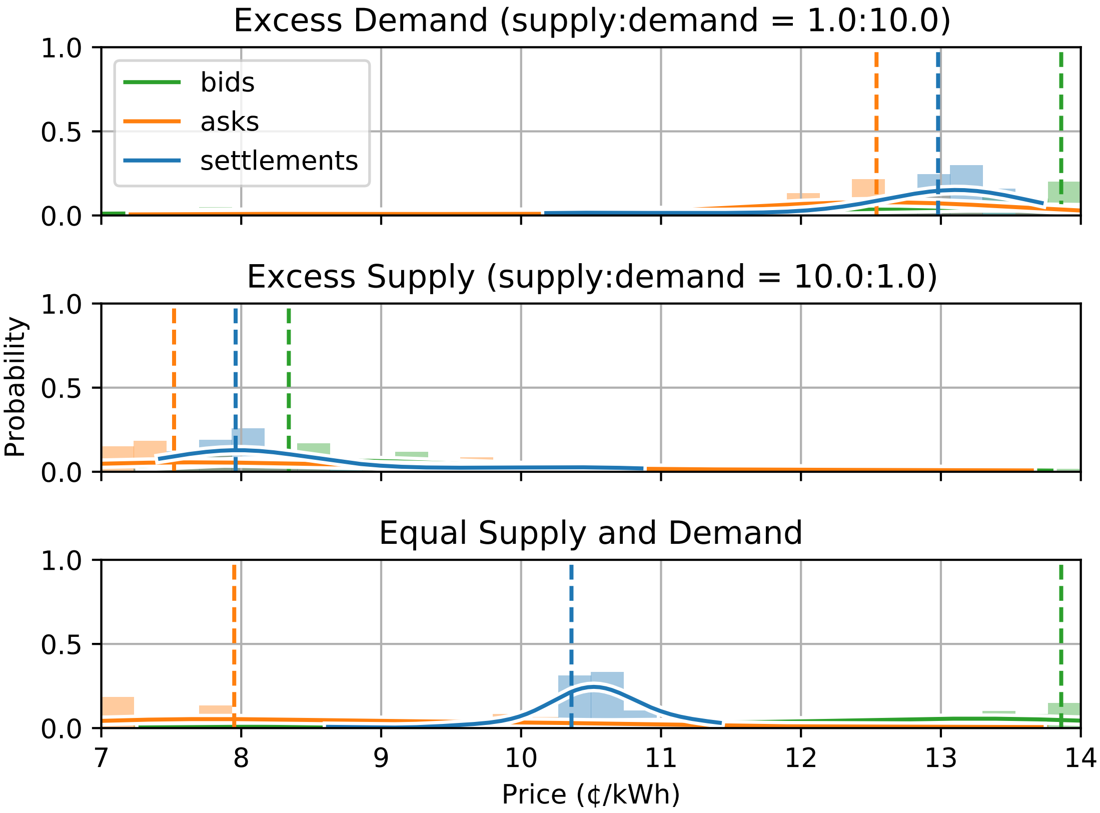

Figure 2.1: Validation policies for bid, ask, and resulting settlement prices for agents operating under M1 for episodes 70 to 100. The histograms show the probabilities of the discrete action policies for the agents and the resulting settlement prices. Modes of the prices are highlighted and shown by the vertical dashed lines. Probability density functions of the histograms are overlaid on top, which are approximated with the Gaussian KDE function in the scikit-learn Python package with default parameters.

Figure 2.2 shows the results for settlement mechanism M2. Unlike the previous mechanism, M2 shows no clear convergence for any case. A possible explanation is that M2 satisfies the conditions for an ideal double auction market, which is known to be impossible to practically implement according to the Myerson–Satterthwaite theorem [62]. Another possibility is that strong budget balancing drastically decreases the number of successful settlements, which results in sparse rewards within this nonstationary environment. Therefore, despite the theoretical existence of a Nash equilibrium, agents are unlikely to discover it due to the lack of feedback. While it may be possible for M2 to converge, given a sufficiently long training time, the fact that it fails to show convergence using the same simulation parameters as M1 and M3 makes it less desirable. RL agents are often expected to update policies on data streams in real-time, which is much slower than in simulations. Currently, there is still a lack of historical smart meter data. Thus, if this system were deployed in a real environment, it would have to learn in real-time. As a result, the market mechanism M2 is disqualified as unsuitable for real world applications.

Figure 2.3 shows the results for settlement mechanism M3. Similar to M1, the bid, ask, and settlement prices are closely clustered on the expected side of the price range for both unbalanced cases, even more closely together. However, unlike M1, the balanced case shows similar behaviour, with prices concentrated in the middle of the price range. Therefore, it can be concluded that M3 has the truthfulness property. The agents have little incentive to set bid/ask prices that deviate too far from the settlement prices, which should closely approximate the true value of energy for each ratio of supply-to-demand. Consequently, M3 is qualified for further research as it satisfies the previously mentioned selection criteria.

The supply-to-demand ratios used in the initial experiments were quite extreme. Examination of the settlement prices for more balanced ratios should provide a more thorough picture of market behaviour. Therefore, the ratios 1.5:1 and 1:1.5 are added to the experiments for M3. The simulations are extended by 100 episodes with annealing as described in Section 2.4.1. As shown by the results in Figure 2.4, the prices settle slightly lower than the balanced case for supply-to-demand of 1.5:1, and slightly higher for supply-to-demand of 1:1.5. This confirms the dominance of the law of supply and demand, as the settlement prices follow the supply-to-demand ratio.

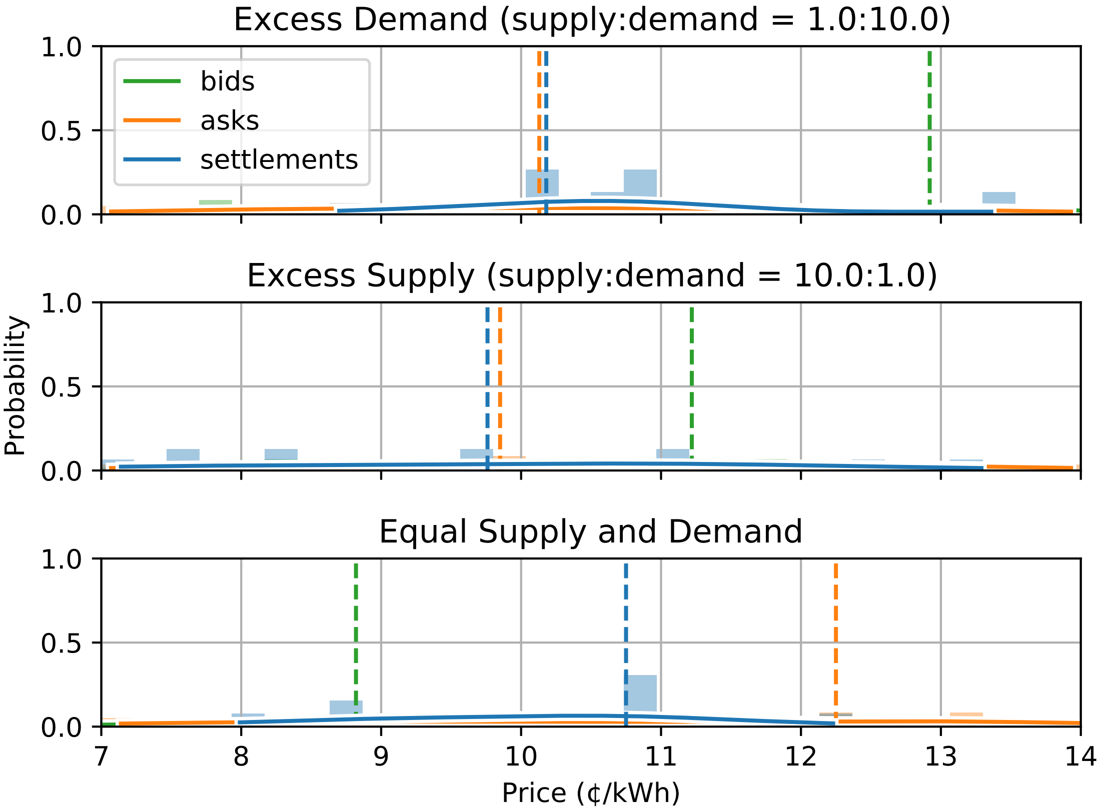

Figure 2.2: Validation policies for bid, ask, and resulting settlement prices for agents operating under M2 for episodes 70 to 100. The histograms show the probabilities of the discrete action policies for the agents and the resulting settlement prices. Modes of the prices are highlighted and shown by the vertical dashed lines. Probability density functions of the histograms are overlaid on top, which are approximated with the Gaussian KDE function in the scikit-learn Python package with default parameters.

Further experiments with more ratios of supply and demand will be performed to develop an empirical model of the price behaviour.

In summary, the experiments show that efficient RL agent training requires weak budget balancing, resulting in a stronger, denser reward signal. Truthfulness is necessary for the emerging policies to truly reflect the law of supply and demand. In a LEM with these two properties, the individual rationality of agents maximizes the value exchanged between participants, guaranteeing economic efficiency as a result of convergence. Even though perfect budget balancing has not been achieved, this may be desirable for deployment: a small profit for the auctioneer can be used to maintain

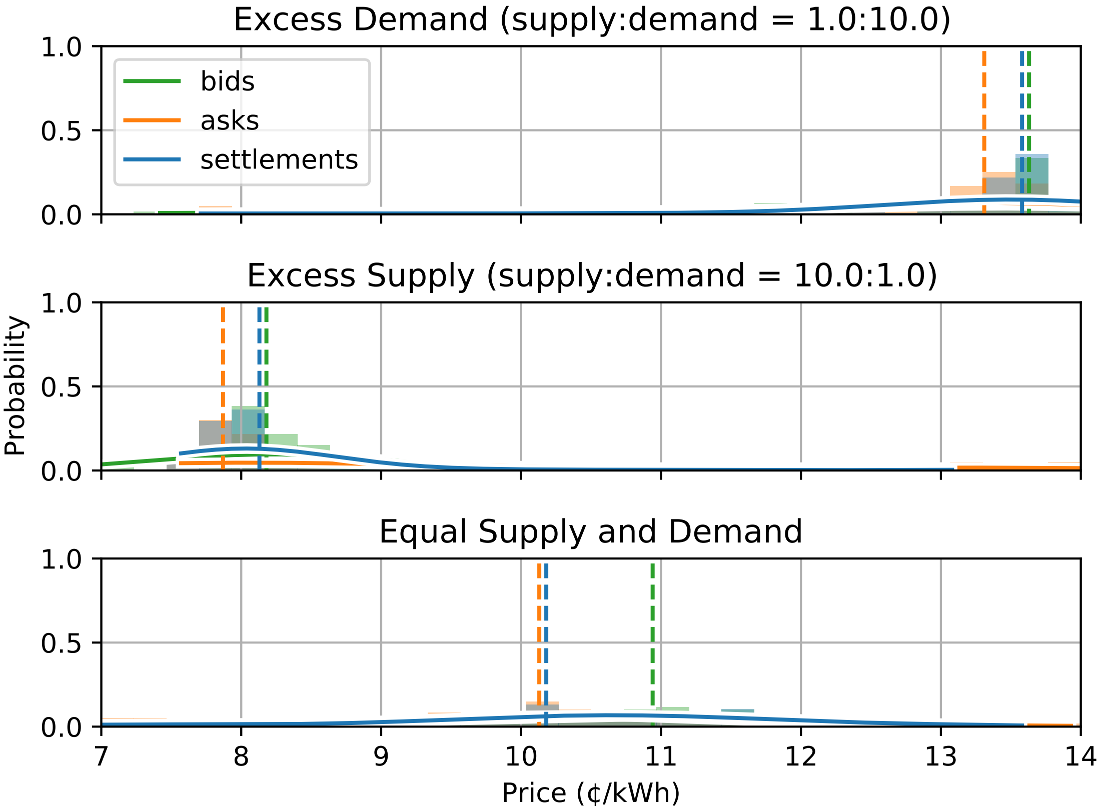

Figure 2.3: Validation policies for bid, ask, and resulting settlement prices for agents operating under M3 for episodes 70 to 100. The histograms show the probabilities of the discrete action policies for the agents and the resulting settlement prices. Modes of the prices are highlighted and shown by the vertical dashed lines. Probability density functions of the histograms are overlaid on top, which are approximated with the Gaussian KDE function in the scikit-learn Python package with default parameters.

the infrastructure necessary for operating the market.

2.4.2 Economic Study

Experimental Design

This experiment focuses on the second research question: Does the resulting market behavior effectively support DR for LEM with high penetration of DER? This investigation is performed by comparing the proposed approach with conventional pricing schemes, such as net billing[5] and time-of-use. As indicated by the results

[5] The distinction between net billing and net metering is clarified in Appendix A.2.

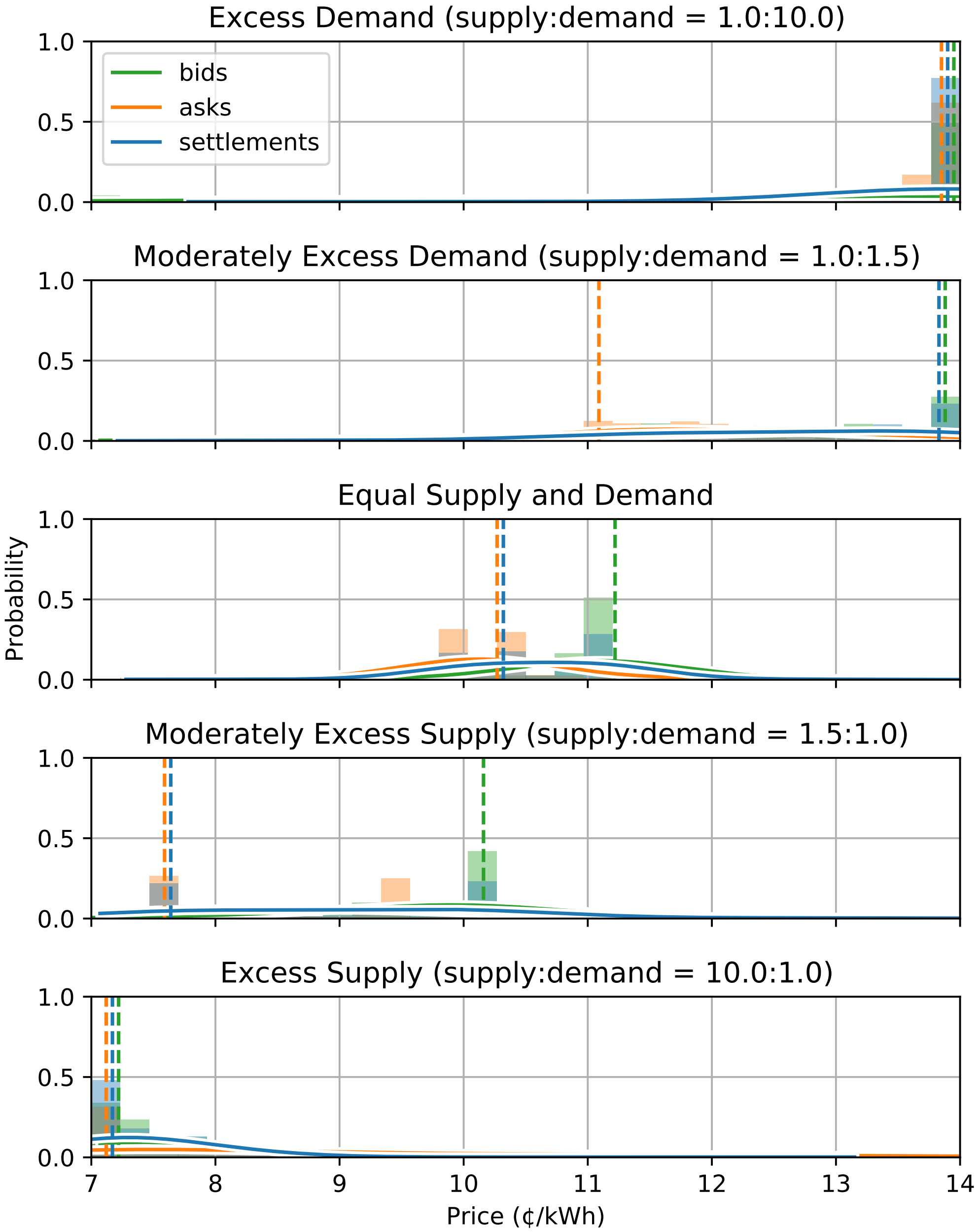

Figure 2.4: Validation policies for bid, ask, and resulting settlement prices for agents operating under M3 for episodes 100 to 200. Exploration factor and learning rate are annealed starting from episode 100 with a multiplier of 0.98 applied at the beginning of each episode. The histograms show the probabilities of the discrete action policies for the agents and the resulting settlement prices. Modes of the prices are highlighted and shown by the vertical dashed lines. Probability density functions of the histograms are overlaid on top, which are approximated with the Gaussian KDE function in the scikit-learn Python package with default parameters.

from the previous experiment, the prices in market equilibrium are dominated by the law of supply and demand (see Figure 2.4). By performing additional simulations with alternative proportions of supply and demand, an empirical model of the price behaviour can be obtained via interpolation. Such a model can be used as a simple approach to set local market prices, without implementing an actual auction-based market. The supply-to-demand ratio for a local market can be derived from metering data.

The economic study is conducted using a residential community microgrid with ten participants. Due to the lack of suitable smart home data from Canada, energy profiles from the openly available SunDance data set [65, 66] are used. Ten energy profiles have been randomly selected to assemble the virtual community. The IDs of the selected customers are as follows: 10011, 1001625, 1002714, 10068, 100703, 1001420, 1003173, 1001230, 100114, 100196. All participants are prosumers participating in energy trading to gain economic benefits. The microgrid is assumed to be on a single bus behind a community smart meter. Similar to the previous experiment, no load shaping is performed. To illustrate the changes of supply-to-demand behaviour of the test community, the aggregated values of supply and demand over a single summer day (June 1, 2015) are plotted in Figure 2.5.

The experiment evaluates the changes of electricity bills caused by enabling a local energy market that sets energy exchange prices based on the local supply and demand. Because the local market price is time-varying and supply-to-demand dependent, the LEM-based pricing can also be compared to a benchmark TOU pricing schedule.

Results and Discussion

The system-wide market model is developed using the data from the previous experiments, supplemented by four additional demand ratios (1:1.1, 1:2, 1.1:1, 2:1). The resulting pricing model is shown in Figure 2.6. Note that in real settings, where the load demand curve and DER availability of each participant are unique, ALEX agents

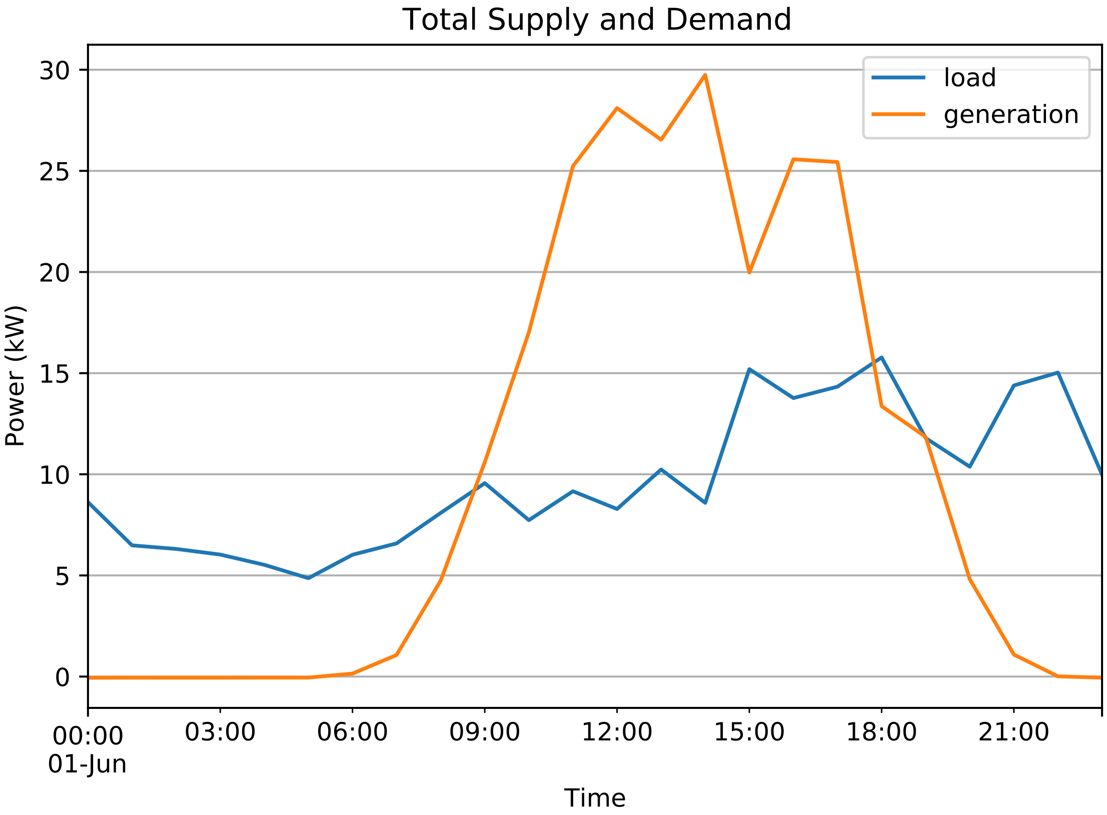

Figure 2.5: Total supply and demand profile of the residential community test over one summer day in June 1, 2015

may develop personalized pricing schedules. The goal of this experiment is to evaluate the economic performance of an ALEX-based trading system and to compare it with common tariffs that do not use individual pricing schedules.

The resulting equation for the price curve is as follows:

$$P(s,d) = \begin{cases} P^{\text{grid}}_{\text{NB,load}} & P(s,d) \geq P^{\text{grid}}_{\text{NB,buy}} \\ P^{\text{grid}}_{\text{NB,gen}} & P(s,d) \leq P^{\text{grid}}_{\text{NB,sell}} \\ -0.0254\frac{s}{d} + 0.1426C^{\text{TOU}}_{H,M,L} \text{ if buying} \\ -0.0280\frac{s}{d} + 0.1299C^{\text{TOU}}_{H,M,L} \text{ if selling,} \end{cases} \tag{2.20}$$

where s is energy supply, d is energy demand, $P^{\text{grid}}_{\text{NB,load}}$ is price of electricity when buying electricity from the grid under net billing, $P^{\text{grid}}_{\text{NB,gen}}$ is price of electricity when selling electricity to the grid under net billing, and $C^{\text{TOU}}_{H,M,L}$ are the adjustment factors when TOU is used.

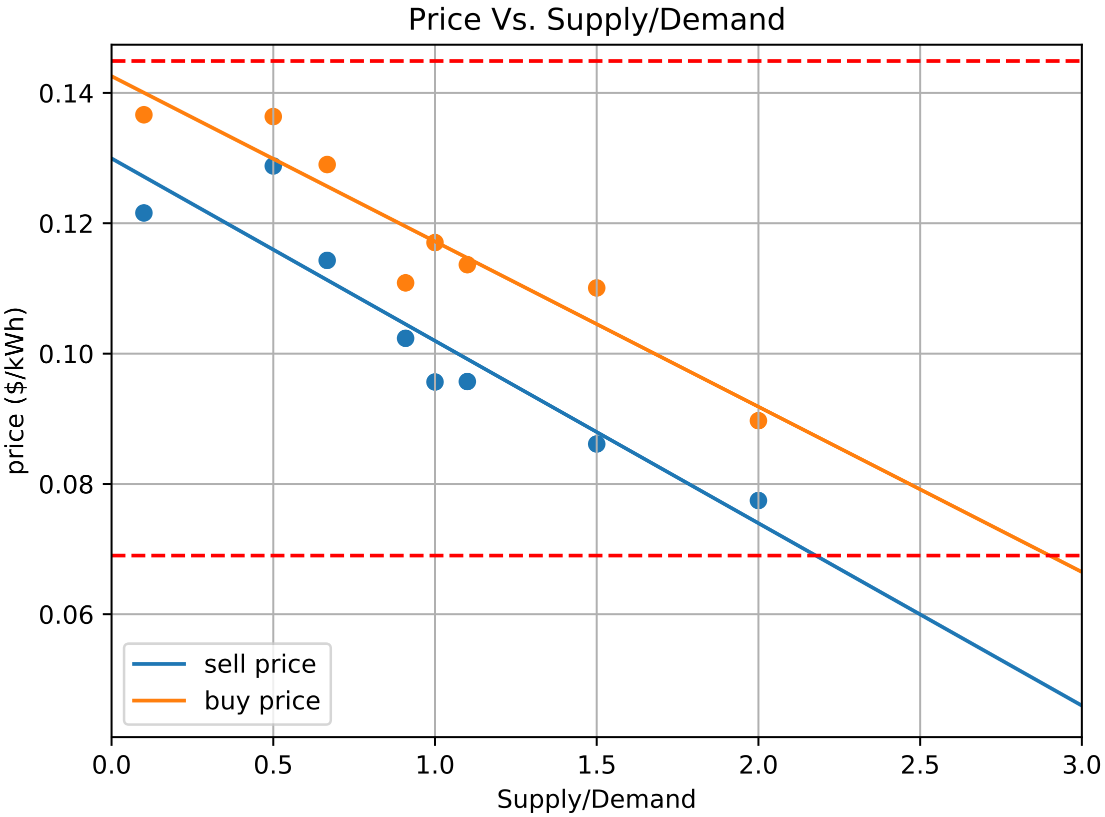

Figure 2.6: Pricing model developed for the test local market. The dotted lines show the price boundaries defined by (2.4). Linear regression between the price points leads to a well-fit, generalized mathematical model.

A pricing schedule for the local market of the energy community and a selected day can be obtained by applying this model to specific energy profiles, as shown in Figure 2.7 for the sample profiles from Figure 2.5.

The internal price determined using the model corresponds well to the ratio of supply and demand of the community throughout the day. For example, at midnight, when solar generation is nil, the price for selling energy to a peer is $0.1449, which is the same for the buyer as if purchasing energy from the grid. Later in the morning, at 7:00AM, when solar energy becomes available, the price for selling to peers lowers accordingly. At around 9:00AM, when generation is significantly higher than demand, the price for selling to peers drops to $0.069, which is the same as selling to the grid under net billing.

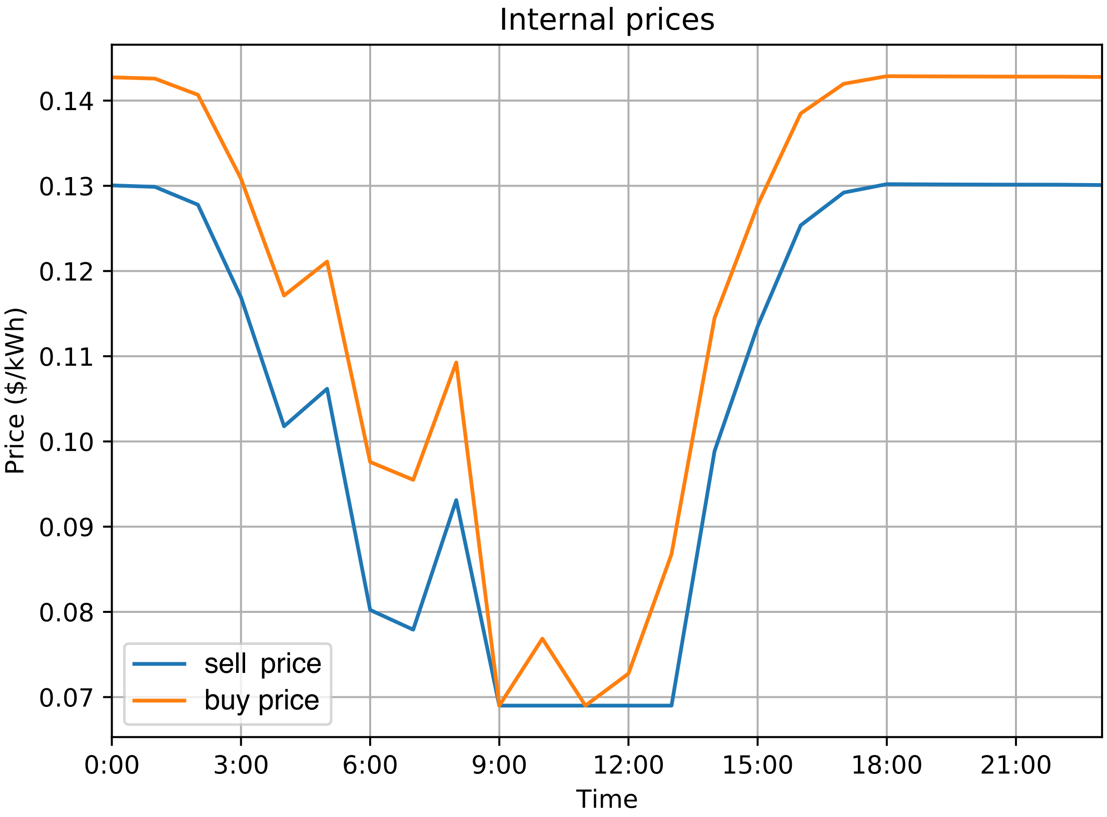

Figure 2.7: Internal prices of ALEX used to conduct transactions

The economic performance of all participants under this price curve is calculated and compared with net billing. The results of this comparison are shown in Figure 2.8 and summarized in Table 2.2. As a reminder, the rules of interaction between the community participants and the grid are detailed in A.4. In accordance with the operating principles of net billing, described in A.2, the entire community is placed behind a community meter, and energy exchanged directly between peers does not incur transmission and distribution fees.

The results in Table 2.2 show that the implementation of a community market can financially benefit the community as a whole, reducing the total community bill by 35.9%. The mean and median of individual bill reductions are 74.51% and 38.8%, respectively. This reduction in bills is due to the more efficient usage of local energy resources, which is more accurately reflected by the local market price. By putting

Table 2.2: ALEX vs. Net Billing (NB)

	Bill ($)			**Energy (kWh) bought from Grid**			**Energy (kWh) bought locally**		**Energy (kWh) sold to Grid**			**Energy (kWh) sold locally**	
Participant	**NB**	**ALEX**	**%**	**NB**	**ALEX**	**%**	**NB**	**ALEX**	**NB**	**ALEX**	**%**	**NB**	**ALEX**
10011	0.10	−0.28	383.2	6.83	2.35	65.59	–	4.48	12.88	0	100	–	12.88
1001625	0.80	0.13	83.2	10.73	2.28	78.75	–	8.45	10.97	0	100	–	10.97
1002714	4.34	3.26	24.8	32.45	13.10	59.63	–	19.35	5.31	0	100	–	5.31
10068	2.80	1.55	44.7	22.37	4.08	81.76	–	18.29	6.41	0	100	–	6.41
100703	2.35	1.32	44.0	22.47	5.94	73.56	–	16.53	13.17	0	100	–	13.17
1001420	2.21	1.47	33.6	17.63	11.96	32.16	–	5.67	4.98	0	100	–	4.98
1003173	7.56	5.09	32.7	61.73	22.00	64.36	–	39.73	20.03	0	100	–	20.03
1001230	3.55	2.69	24.2	24.89	8.03	67.74	–	16.86	0.80	0	100	–	0.80
100114	5.22	2.70	48.2	49.55	12.65	74.47	–	36.9	28.40	0	100	–	28.40
1001965	6.50	4.78	26.5	50.78	19.07	62.45	–	31.71	12.45	0	100	–	12.45
Total	**35.43**	**22.70**	**35.9**	**299.44**	**101.46**	**66.12**		**198.00**	**115.40**	**0**	**100**		**115.40**

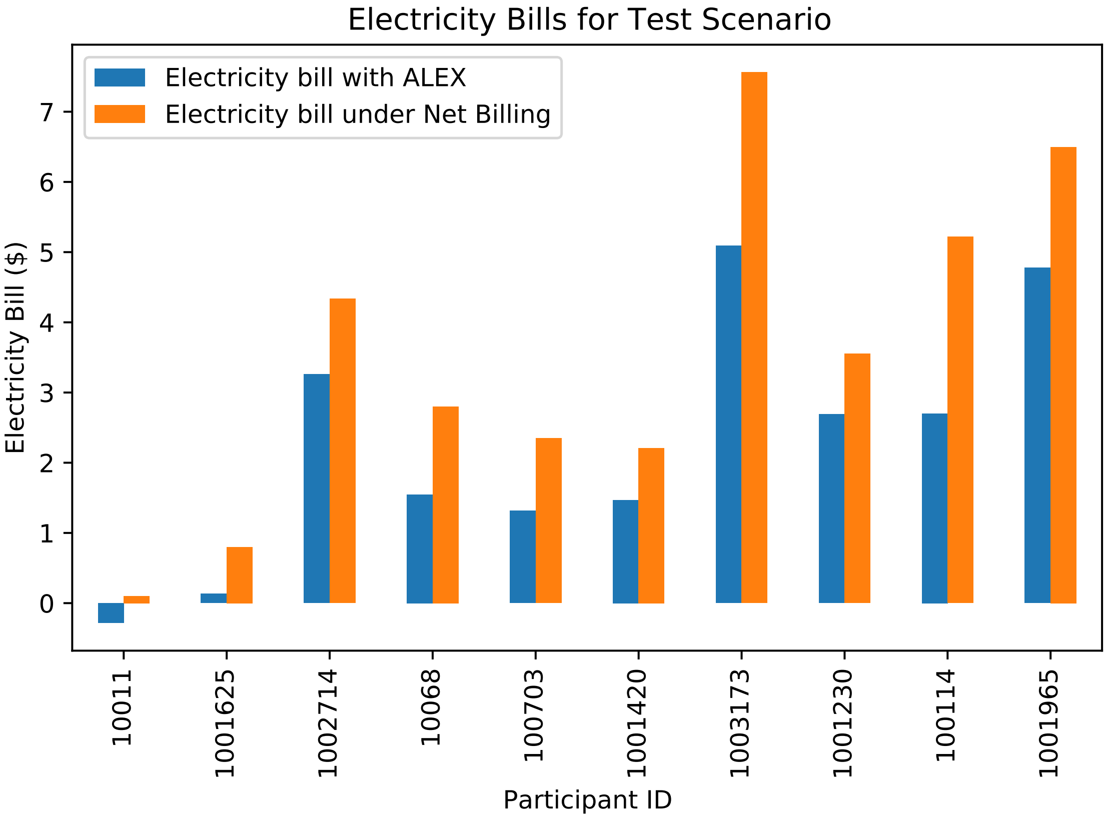

Figure 2.8: Electricity bill comparison between net billing and ALEX

the whole community behind-the-meter, financial benefits may even be gained by those who cannot afford the expense of acquiring and installing their own DER. This is because they have direct access to excess generation of their neighbours that may be lower priced in comparison to buying from the grid. Similarly, there is an inherent financial incentive to sell excess generation to peers first, as the profits can be higher than selling directly to the grid. In other words, this setup may further socialize the benefits of DERs. Although not directly comparable, recent work by Jogunola et al. [55] obtained average financial benefits of 35% by leveraging energy storage, and 55% when leveraging both PV and storage. While the proposed approach currently uses only PV, energy storage will be added in future studies. Similar or better performance than that reported in [55] is expected.

As mentioned before, the local market price is time-varying and supply-to-demand dependent. This is similar to the philosophy behind the development of TOU, which uses the time-varying supply/demand of the entire grid instead of focusing on any specific area. Therefore, these two approaches are compared to quantify their relative performance. Ontario TOU is used as a benchmark in Canada and is often referenced by utility companies in jurisdictions without TOU, such as Alberta.

Figure 2.9 displays the two pricing schedules, showing the stark contrast between their shapes. Whereas the local energy price decreases toward noon due to the increase in generation, TOU increases, which suggests that there is more load than generation during this period. While it is possible that this is true due to commercial and industrial loads, which do not exist in the local market, the fact remains that the TOU is not correlated with the actual balance of load and demand in the testing locale. While this disconnect may be a cause for the lack of participation in TOU mentioned in Section 2.2, it also suggests the need for very localized, highly relevant pricing signals to increase the efficiency of managing DERs and the overall system. ALEX is a highly scalable, distributed approach that generates highly relevant pricing signals at a low cost.

2.5 Conclusions and Future Work

DR techniques provide an effective means to manage DERs. As more such resources are installed and the energy mix becomes more complex, their management and coordination should be automated. This article explores the requirements for automating LEMs using multi-agent reinforcement learning. The exploration is facilitated by ALEX, a LEM framework that can use an arbitrary closed-book, double auction settlement system. It is used to identify the market properties that drive the policies of independent Q-learning agents to follow the law of supply and demand. After establishing an appropriate market settlement mechanism, the emergent market behaviour is compared to conventional DER integration techniques.

The first experiment trains a group of agents with three market configurations,

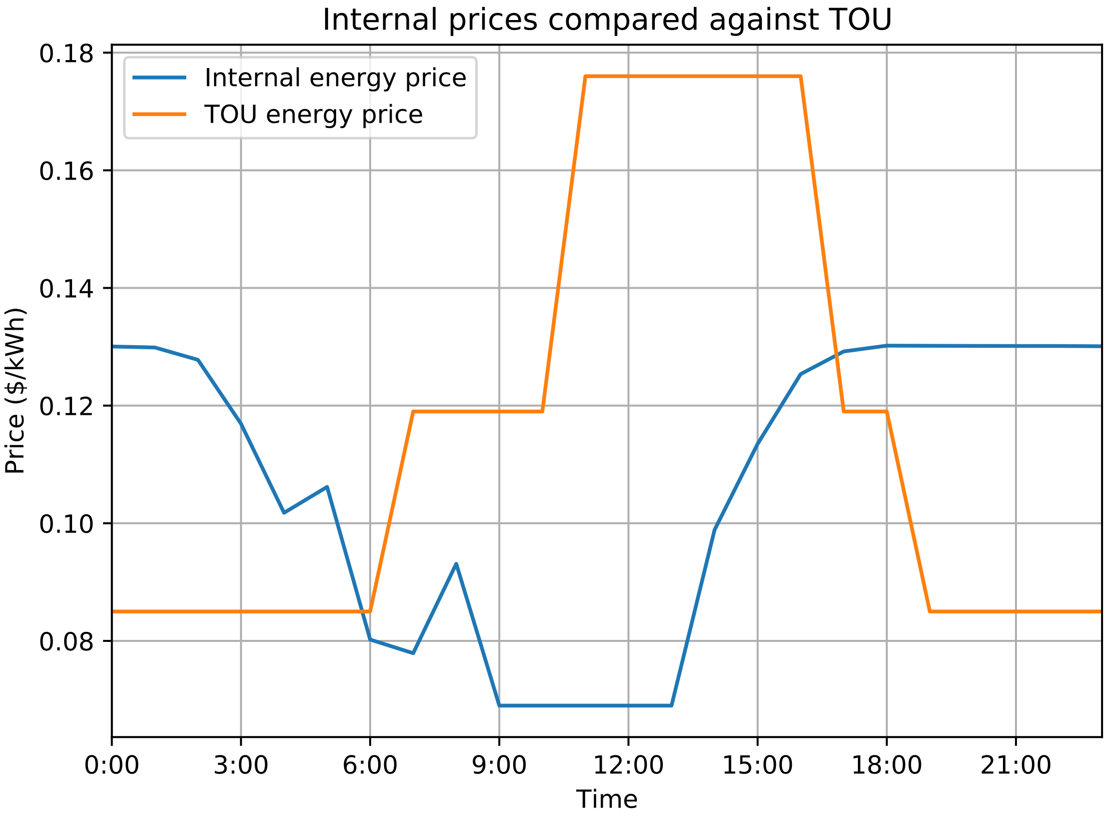

Figure 2.9: Local market pricing schedule compared against Ontario summer TOU prices.

distinguished by their general properties. The results show that, for a double auctions based market, truthfulness is necessary for the collective policy to reflect the law of supply and demand. The second requirement, weak budget balancing, facilitates the generation of stronger, denser reward signals sufficient for training the trading agents. These properties, combined with agents' individual rationality, maximizes economic efficiency and reduces the effects of weak budget balancing.

The second experiment compares the resulting LEM behaviour with that of markets based on net billing and time-of-use (TOU). Since consensus pricing in ALEX strongly reflects the law of supply and demand, the resulting price signal is significantly more responsive and relevant than TOU. This signal will likely further increase the effectiveness of DER utilization. In turn, these efficiency gains will propagate upstream across the entire grid. One such effect can be observed in the economic performance of the test community. Using the proposed approach, the community as a whole experienced a bill reduction of 35.9% compared to net billing. For individual customers, the mean and median bill reductions were 74.51% and 38.8% respectively.

The findings presented in this article lay the crucial ground for future work. We plan to investigate the integration of battery energy storage systems, along with more complex RL algorithms to increase profile shaping capabilities. These extensions are expected to further reduce bills, as well as to increase the grid efficiency and stability. Performance comparisons with other deployed LEM approaches will be conducted to identify the most suitable automation approach for ALEX. As the development of this new LEM framework continues, additional studies will examine the effects of its application. One planned study is the evaluation of system efficiency improvements within a local market with the addition of battery storage and electric vehicles. This will be further extended to study the impact across the grid via inter-ALEX energy exchanges.

Chapter 3

Transactive Local Energy Markets Enable Community-Level Resource Coordination Using Individual Rewards

3.1 Introduction

The electrification and the accelerated adoption of distributed energy resources (DERs) significantly alter community net load patterns, exacerbating variability and giving rise to phenomena like the "duck curve" [4]. The uneven distribution of DERs, coupled with the challenges they pose, results in localized grid disturbances. Traditional stabilization methods, such as spinning, primary, and secondary reserves, lack the required local granularity. Demand response (DR), which uses control signals to shape the community net load at the building level, has emerged as a promising and effective solution to address these issues.

DR can be classified as direct or indirect, depending on the nature of the control signal. Direct DR enables the grid operator to shape the net load by directly accessing controllable assets. However, its adoption faces challenges due to the inherent conflict of interest between the grid operator and end users regarding asset usage. On the other hand, indirect DR aims to encourage more favorable net load patterns through a proxy control signal, often utilizing monetary incentives. Although indirect DR is easier to convey to end users, it is often found to be ineffective [9].

According to the GridWise Architecture Council [10], transactive energy (TE) is defined as "the use of a combination of economic and control techniques to improve grid reliability and efficiency." A TE system falls under indirect DR, intending to balance its supply and demand in a decentralized (autonomous) fashion through well-aligned incentives. The local energy market (LEM) has recently emerged as a framework to implement TE at the grid-edge. Mengelkamp et al. [20] define LEM as "a geographically distinct and socially close community of residential prosumers and consumers who can trade locally produced electricity within their community. For this, all actors must have access to a local market platform on which (buy) bids and (ask) offers for local electricity are matched."

Therefore, the LEM functions as an indirect and decentralized DR mechanism, with the aim of aligning the goals of the grid stakeholders and participating electricity end users through dynamic electricity pricing that reflects the state of the grid and the community. Under this assumption, the stability and efficiency of the local electricity grid improve as participants minimize their electricity bills and maximize their DER-related returns. Consequently, much of the LEM literature evaluates market performance based on economic metrics, presuming that minimizing bills is equivalent to positive effects on the local electricity grid [19].

An emerging body of literature argues that the underlying assumption of incentive alignment does not automatically hold true [67–69]. LEM studies are inherently scenario-based and performance depends both on participant behavior and the market mechanism of the EM [20, 67]. As noted by Mengelkamp et al. [20] and reinforced by contributions from Kiedanski et al. [68] and Papadaskalopoulos et al. [69] the complex interplay between strategies and market design has a profound impact on LEM performance. A more thorough approach is needed to adequately assess the efficacy of a given LEM design as a DR system.

Following this logic, this study focuses on the DR capabilities of the Autonomous

Local Energy eXchange (ALEX), originally proposed by Zhang et al. [1]. It investigates ALEX's ability to foster the desired alignment of interest between participants and grid stakeholders, and further explores ALEX's capabilities as a community-level DR system.

Assuming selfish behavior, near-optimal participant policies are generated using a dynamic programming-based procedure, and their performance is evaluated on an open-source dataset. The results demonstrate that within ALEX, selfish cost minimization leads to the emergence of community-level DER coordination, significantly improving several metrics related to community net-load, such as ramping rate, load factor, and peak load. They exhibit community-level coordination of DERs, facilitated by the LEM. Although participants only have access to building-level information, they clearly outperform classical indirect DR approaches that operate at the building level. This study also outlines a methodology for evaluating LEM based on a high-quality open-source data set previously used to analyze other DR systems [70, 71]. This contribution thereby helps alleviate the lack of benchmarks in this field.

The article is organized into seven sections. Related work and background concepts are described in Section 3.2. Section 3.3 deduces the investigated hypotheses and develops the experiments, algorithmic methodology and metrics to evaluate them. Experimental results are described and discussed in Section 3.4. The final section provides a brief summary of the study, draws the main conclusions, and outlines possible directions for future work.

3.2 Background

This section reviews literature related to this contribution regarding methodology, data sets, and evaluation metrics. It also provides background information on ALEX and markov decision processes.

3.2.1 Related Literature

Capper et al. [19] and Mengelkamp et al. [20] offer a comprehensive review of recent LEM literature. The authors categorize LEM approaches as peer-to-peer, individual, community-level self-consumption, and TE LEM. The majority of studies focus on the economic performance of end users within the LEM compared to net billing. For example, Mengelkamp et al. [16, 72] compare several LEM designs using a range of heuristics to demonstrate improved economic performance of the proposed LEM.

Kiedanski et al. [68] and Papadaskalopoulos et al. [69] study LEM scenarios in which agent-to-market interactions lead to detrimental effects on the local electricity grid. This can occur due to a misaligned incentive signal conveyed through a suboptimally designed LEM or due to ill-tuned, suboptimal policies resulting in participant actions that deviate from the incentivized policy. This investigation serves as one of the primary objectives of our study: exploring ALEX's capabilities to align participant and grid stakeholder incentives. We approach this by generating a set of near-optimal policies, using an appropriate search algorithm rather than relying on heuristics to assess ALEX's performance.

A deeper understanding of LEMs as DR systems requires investigating their effects on the local electricity grid. A review of LEM by Dudjak et al. [67] focuses on the impacts of LEMs on power systems. The authors highlight a challenge underlying the direct comparison of LEM articles that rely on power-flow-related metrics, such as voltage violations and congestion: Direct power-flow analysis injects design choices, such as circuit and load placement, into the experiment, exacerbating the problem of LEM comparability, as there are no clearly adopted benchmark scenarios across the community. To avoid this issue, this article performs an experimental analysis using metrics that pertain to power system stability and efficency, but are circuit-independent. It relies on net load-related metrics such as ramping rate, load factor, peak export, and import to provide insight into variability across several time-scales.

In addition, to ensure robust results, the experiments utilize an openly accessible, high-quality data set of sufficient length.

Nweye et al. [71] and Vázquez-Canteli et al. [73] describe community-level DR approaches based on deep reinforcement learning (DRL), a concept that is related to the simulation approach employed in this article. Refer to Vázquez-Canteli et al. [27] for a comprehensive review of articles that apply reinforcement learning to DR. Nweye et al. [71] conduct their DR experiments using the same data set and a set of metrics similar to those in this study. Nevertheless, ALEX relies on building-level information and optimizes for a singular, building-level objective, whereas Nweye et al. [71] use both building- and community-level information for control and optimize for a mixed objective that includes metrics at both levels. Although this study does not directly compare with [71], we maintain principal comparability by using the same data set and similar metrics.

Zang et al. [49] investigate a decentralized peer-to-peer LEM, where buildings within a community and a centralized battery energy storage system (BESS) supply bids and asks to the LEM. The BESS is automated via an RL agent that learns to maximize its profit through temporal arbitrage and load shifting. As mentioned earlier, the current article uses grid performance-related metrics to evaluate the DR capabilities of ALEX. In contrast, study [49] uses only economic metrics to evaluate LEM performance and features a single rational agent with a monopoly on the load-shifting service. The LEM introduced in this article has several independent buildings, each represented by a rational agent equipped with load-shifting capability.

Xu et al. [74] propose a methodology for DR using community-level dynamic pricing. They employ neural networks to forecast load consumption and subsequently develop an approximate pricing schedule. The forecasts and schedules are then used to formulate a markov decision process (MDP), which is approximately solved using Q-learning. Although [74] and this study employ a conceptually similar simulation approach, there are several distinguishing factors. Both studies evaluate the performance of pricing mechanisms inherently tied to the ratio of supply and demand. However, in Xu et al.'s [74] study, they employ a schedule-based dynamic pricing model that remains fixed during the optimization process, even when the balance between supply and demand changes. In ALEX, the equilibrium price reflects market dynamics and thus changes due to actions performed by agents. Both studies simulate DR methods by formulating and subsequently solving MDPs. However, the current study uses a fundamentally different method to simulate agents, based on a tree-search approach with strong convergence properties.

3.2.2 Autonomous Local Energy Exchange (ALEX)

ALEX is a purely economy-driven LEM, where the price of a specific energy transaction is not dictated by its impact on one or several metrics of local electricity grid performance. Instead, it results from participants' efforts to minimize their bills. We reframe ALEX, initially proposed by Zhang et al. [1] as an automated, economy-driven LEM, using the nomenclature established by Chapper et al. [19]: ALEX is an LEM that facilitates trading between buildings b of a community B through a blind double auction settlement mechanism based on clocks and futures. Clock-based markets employ bids and asks supplied for specific settlement time steps, rather than following a continuous settlement process. The futures market means that at the current time step t_{now}, bids and asks are accepted for a future settlement time step $t_{\text{settle}} > t_{\text{now}}$, and the settlements are delivered at a subsequent time step $t_{\text{deliver}} > t_{\text{settle}}$. Blind double action means that each building b communicates bids and asks to the market without seeing bids and asks of other buildings.

Using only building-level information, each building b minimizes its electricity bill calculated as

$$\begin{aligned} bill_b = & \left(cost_b^{\text{market}} - profit_b^{\text{market}}\right) + \\ & + \left(cost_b^{\text{grid}} - profit_b^{\text{grid}}\right) + fees_b, \end{aligned} \tag{3.1}$$

that is as the sum of the market bill ($cost_b^{\text{market}} - profit_b^{\text{market}}$) and grid bill ($cost_b^{\text{grid}} - profit_b^{\text{grid}}$), in addition to the fee component $fees_b$.

The utilization of LEM is incentivized through a profitability gap

$$p^{\text{grid,sell}} < p^{\text{market,min}} <= p^{\text{market}} <= p^{\text{market,max}} < p^{\text{grid,buy}}. \tag{3.2}$$

This profitability gap enables a mutually advantageous exchange of energy on the LEM at a market price p^{market}, ranging between the minimum market price $p^{\text{market,min}}$ and the maximum market price $p^{\text{market,max}}$. The minimum and maximum market prices are constrained by the grid sell price $p^{\text{grid,sell}}$ and the grid buy price $p_{\text{grid,buy}}$, respectively. A profitability gap can be achieved through various mechanisms, for example, GHG or fee offsets [19, 67]. The size of the profitability gap has no impact on the hypotheses investigated and the metrics used in this study.

Optimal actors within ALEX converge to a Nash equilibrium due to its nature as a (partially observable) stochastic game [1]. The authors conducted an in-depth investigation into ALEX's settlement mechanism [1], identifying a design in which a set of agents learns to price in relation to the supply/demand ratio, despite having no information about it. These experiments were carried out without the presence of load-shifting capacity. The follow-up study [75] evaluates a system with one residential battery controlled by an expert-designed heuristic.

This article builds on these studies and significantly extends their contribution by investigating ALEX's aligning properties and its capabilities as a DR system. The analysis of ALEX's DR properties is conducted in the presence of load-shifting capabilities, using rational agent behaviors. A rational agent aims to perform optimally with respect to its environment and objective.

3.2.3 Markov Decision Processes

An MDP is defined as a tuple (S, A, P_a, R_a), which forms a model of a discrete-time stochastic control process [76]. The process nodes, or states, are fully described through the state space S, the set of all possible states $s \in S$. The action space A is the set of all possible actions $a \in A$. An action a initiates the transition from the current state s to the next state s' with a transition probability $P_a(s, s') \in [0, 1]$. This transition results in a reward $r = R(s, s')$. The transition probability can also be denoted as $p(s', r|s, a)$, specifying the transition from s to s' using a while receiving a reward r. The optimization objective of a MDP is the return G, defined as the discounted cumulative sum of future rewards, given the discount factor $\gamma \in [0, 1]$ and a sequence of state transitions

$$G_t = \sum_{i=t}^{\infty} \gamma^{i-t} R\big(s^i, s^{i+1}\big). \tag{3.3}$$

A policy, π, is a (probabilistic) mapping $S \mapsto A$. It allows to define the state value $V(s)$ as the expected return G, given a policy π followed from a starting state s

$$V(s) = \mathbb{E}\left[G|s, \pi\right]. \tag{3.4}$$

The optimal policy, π^*, maximizes V

$$V_{\pi^*}(s) = \max_{\pi} V_{\pi}(s). \tag{3.5}$$

There are several common search methods for MDPs to determine the optimal policy π^*, including dynamic programming [48, 77], Monte Carlo tree search [78], and reinforcement learning [48]. This study employs dynamic programming using value iteration as described by Sutton et al. [48]. The pseudocode of this method is outlined in Algorithm 1.

3.3 Methodology

The goal of this study is to investigate ALEX's DR capabilities in the presence of multiple independent agents with the capability to load-shift. Some LEMs ensure alignment between participant and grid stakeholder interests by tying pricing directly to grid stability or related metrics, whereas ALEX is a purely economy-driven LEM.

Algorithm 1 Dynamic programming through value iteration, as per Sutton et al. [48]. *tol* is the convergence tolerance. V is the state value (3.4). $p(s', r|s, a)$ is the probability of the transition to the next state s' while receiving the reward r, starting in state s given action a. π is the policy. γ is the discount factor.

```
given MDP
given $tol, \gamma$
$\delta = \infty$
while $\delta > tol$ do
    $\delta = 0$
    for $s \in S$ do
        $V_{\text{old}}(s) = V(s)$
        $V(s) \longleftarrow \max_a \sum_{s',r} p(s', r|s, a) \left[r + \gamma V(s')\right]$
        $\delta = \max(\delta, |V(s) - V_{\text{old}}(s)|$
    end for
end while
Output deterministic policy $\pi \approx \pi^*$ such that
$\pi(s) = \arg\max_a \sum_{s',r} p(s', r|s, a) \left[r + \gamma V(s')\right]$
```

3.3.1 Study Hypotheses

In light of recent studies questioning the alignment capabilities of economy-driven LEMs, we investigate the following hypothesis:

Hypothesis 1: ALEX's market mechanism, which incentivizes bid and ask prices in correlation to the current timestep's supply and demand ratio within a profitability gap, is capable of fostering a strong alignment between participant and grid stakeholder interests;

This is based on intuition revolving around the competitive nature of ALEX. Rational agents should compete for the most profitable arbitrage opportunities, striving to maximize their own bill savings. Concurrently, they utilize the load-shifting capacity to manipulate the supply/demand ratio in their favor. ALEX's market mechanism should be sufficient to strongly encourage interest alignment.

It is economically rational to maintain local surplus generation within the community. Shifting surplus generation through the LEM is more profitable than selling it to the grid and subsequently satisfying the load demand from the grid (3.2). Each agent would first aim to meet its load demand through LEM. Supplying to the market during times of high demand is more profitable than doing so during times of low demand, and purchasing from the market during times of high supply is more profitable than purchasing during times of low supply. This informs the deduction of a second hypothesis:

Hypothesis 2: Rational agents, representing individual buildings within ALEX, coordinate DER usage patterns across the entire community B, despite each agent operating with only building-level information and selfishly minimizing its own electricity bill.

Agents within ALEX do not share information. Communication and information sharing, common in other LEM designs, incentivize community-level coordination [19]. Demonstrating hypothesis 2 would illustrate a set of unexpected yet desirable properties for ALEX. The rational agents within ALEX converge to a Nash equilibrium, meaning that their policies are best responses to each other and each agent is maximally exploiting the joint communal policy. The presence of several agents with load-shifting capacity should stabilize the LEM's supply/demand ratio through temporal arbitrage in the market. This results in load-flattening behavior at the community level across both short- and long-term time scales, resembling the properties of centralized DR systems. It has a higher performance ceiling than building-level DR, despite each agent only using building-level information.

Demonstrating both hypotheses would highlight ALEX as an efficient tool for implementing TE at the community level. This approach allows for the realization of community-wide benefits without the necessity of centralization or data-sharing.

3.3.2 ALEX as a Markov Decision Process

The joint state-space of ALEX, S_B, covers the entire community B and is the product of the individual building state spaces, S_b, of all buildings $b \in B$

$$S_B = \prod_{b \in B} S_b. \tag{3.6}$$

The state of an individual building, s_b, is as a tuple of the load demand $l(t)$, generation $g(t)$ for time step t, and state of charge of the battery SoC

$$s_b = (l(t), g(t), SoC) \tag{3.7}$$

This notation can be condensed into a tuple of time step t and SoC

$$s_b = (t, SoC). \tag{3.8}$$

The joint action space A_B is the product of the individual action spaces A_b of all buildings b

$$A_B = \prod_{b \in B} A_b. \tag{3.9}$$

The actions of individual buildings, a_b, are tuples of bid bid_b, ask ask_b, and battery action bat_b

$$a_b = (bid_b, ask_b, bat_b). \tag{3.10}$$

The bid bid_b and the ask ask_b are tuples consisting of price p and quantity q, respectively, according to

$$bid_b = (p_b^{\text{bid}}, q_b^{\text{bid}}), \tag{3.11}$$

$$ask_b = (p_b^{\text{ask}}, q_b^{\text{ask}}). \tag{3.12}$$

The joint MDP of ALEX is deterministic, with a transition probability equal to 1. Since the goal of each agent is to minimize the bill $bill_b$ of its building b, defined by (3.1), the reward function R_b is the negative of the bill

$$R_b(s, s') = -bill_b(s, s'). \tag{3.13}$$

The joint community policy of ALEX, π_B, is the set of individual building policies π_b.

With ALEX defined as an MDP, a search method can be developed to find a near-optimal policy $\pi_b \approx \pi_b^*$ for rational agents.

3.3.3 Simulation of Rational Agents for ALEX

The primary focus of this study is to examine the system properties of ALEX. In contrast to RL-based approaches (such as Xu et al. [74]), this study uses dynamic programming through value iteration to search the MDP for an optimal set of deterministic building policies π_B^*. Dynamic programming, in comparison to Deep Reinforcement Learning (DRL) algorithms designed for the same setting, exhibits robust convergence properties, rendering it better suited for this particular task. While generating a set of generalizing agents does require a learning approach (such as DRL), the evaluation of these agents' capabilities becomes challenging without a well-founded understanding of ALEX's performance potential. Therefore, we defer the analysis of DRL agents to future work, focusing the contribution of this study on establishing an in-depth understanding of ALEX's systemic properties.

While it is possible to use dynamic programming to search the joint state space S_B and the joint action space A_B for the optimal joint policy, π_B^*, such a process would be extremely time-consuming. To enhance the efficiency of this search, several adjustments have been made, as outlined in the remainder of this section.

Zhang et al. [1] show that the optimal communal policy, π_B^*, is expressed as the Nash equilibrium of individual building policies π_b^*, where each building policy is the best response to all other policies that currently compose π_B^*. Therefore, the approach used in this study iteratively computes the best response of each building π_b^* to the current communal policy π_B, randomly iterating through buildings $b \in B$. This way, only one building policy π_b changes at a time, maintaining convergence to a Nash equilibrium while searching a significantly smaller space. The search is performed in the building state space S_b and the building action space A_b for each building in the community $b \in B$. This approach effectively replaces $\prod_{b \in B}$ with $\sum_{b \in B}$, i.e.

$$\sum_{b \in B} S_b < S_B = \prod_{b \in B} S_b, \tag{3.14}$$

$$\sum_{b \in B} A_b < A_B = \prod_{b \in B} A_b. \tag{3.15}$$

In addition, the building action space A_b is simplified using the price curve derived by Zhang et al. [1] as a rational heuristic for the bid price p_{bid} and the ask price p_{ask}.

The net load of each building at time step t, denoted as $E_b(t)$, is defined as the sum of load demand $l(t)$, generation $g(t)$, and battery charge $bat(t)$:

$$E_b(t) = l_b(t) - g_b(t) + bat_b(t). \tag{3.16}$$

The bid and ask quantities q_{bid} and q_{ask} are set as the residual positive and negative net load, respectively. The battery charge bat_b gives each agent the ability to manipulate E_b and, consequently, the market interactions.

$$q_{\text{bid}}(t) = \max(E_b(t), 0), \tag{3.17}$$

$$q_{\text{ask}}(t) = \max(-E_b(t), 0). \tag{3.18}$$

The state of charge of each building is discretized into n_{quant} discrete values to iterate over each state s of the building state space S_b. This discretization results in a manageable size for the building state, given by:

$$|S_b| = T n_{\text{quant}}, \tag{3.19}$$

where b represents the building and T is the number of time steps. In the experiments, n_{quant} is set to 40.

The building action space A_b is quantized to align with S_b by only allowing bat_b transitions from one valid state to another, denoted as $s_b, s'_b \in S_b$. This restriction limits the number of actions for any state to a maximum of n_{quant}. This forced quantization, combined with the search for deterministic policies, may lead to situations where reaching π_B^* is unattainable, leading to cyclic sequences of policies that revolve around the true Nash equilibrium. To address this, a policy distance-based cut-off criterion is introduced based on the mean, state-wise difference

$$d_{\pi_{\text{new}},\pi_{\text{old}}} = \frac{1}{|S_b|} \sum_{s \in S_b} |\pi_{\text{new}}(s) - \pi_{\text{old}}(s)|. \tag{3.20}$$

When the distance metric for each building remains below 0.01, the joint community policies are considered converged, i.e., $\pi_B \approx \pi_B^*$. This arbitrarily chosen threshold effectively avoids cyclical convergence patterns in this study. The abstract pseudocode of the algorithm is provided in Algorithm 2 and implemented in Python.

Algorithm 2 A pseudocode of the algorithm to generate rational agents for ALEX, given the MDP discussed in Section 3.3.2. DP refers to the dynamic programming algorithm defined in Algorithm 1, and D is the distance metric defined in Formula 3.20.

```
given MDP
for $b \in B$ do
    initialize $\pi_b$ randomly
end for
$d_B = \infty$
while $d_B < 0.01$ do
shuffle $B$
   for $b \in B$ do
    $\pi_b^{\text{old}} = \pi_b$
    $\pi_b^* = \text{DP}(S_b)$
    $\pi_b = \pi_b^*$
    $d_b = D(\pi_b^*, \pi_b^{\text{old}})$
   end for
$d_B = \max_{b \in B} d_b$
end while
```

3.3.4 Evaluation Methodology

This section discusses the design of a set of experiments to test the hypotheses presented in Section 3.3.1. The hypotheses are evaluated on the CityLearn2022 data set [70], which provides a year of hourly data for 17 smart community buildings in an open-source format. For each building, it includes a time series of energy demand (l_b) and photovoltaic generation (g_b), along with details of the BESS. The open-source nature of this dataset enables follow-up studies to benchmark directly against this contribution, spanning a variety of DR applications.

To maintain comparability with other studies using this dataset, and due to the absence of available benchmark circuits, ALEX's performance is assessed using a set of community net load metrics. This approach provides insights into the variability of net community load across various time scales, which is generally relevant to power system stability. To maintain primary comparability with previous studies on the CityLearn2022 data set, such as Nweye et al. [71], we adopt and extend previously used metrics. Economic metrics, such as carbon emission rate, electricity price, and economic welfare, are excluded as they are not directly related to the hypotheses examined in this study. Nevertheless, for a general comparison across literature, an overview of average electricity bills can be found in the Appendix B.1.

All performance metrics in this study are functions of the community net load E_B, calculated as the sum of the net loads of all buildings

$$E_B(t) = \sum_{b \in B} E_b(t). \tag{3.21}$$

In the following expressions, n_d is the number of days in the data set, d is the number of time steps in a day, and t is the current time step. $\max_{\text{start}}^{\text{stop}}$ and $\min_{\text{start}}^{\text{stop}}$ denote, respectively, the maximum and minimum values over the interval from start to stop. Given the hourly resolution of the CityLearn2022 data set, the conversion from kWh to kW is straightforward and, therefore, is excluded from the notation.

The average daily imported energy

$$\overline{E}_{d,+} = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(\sum_{t \in d} \max(E_B(t), 0) \right), \tag{3.22}$$

and the average exported energy

$$\overline{E}_{d,-} = \frac{-1}{n_d} \sum_{d=0}^{n_d} \left(\sum_{t \in d} \min(E_B(t), 0) \right), \tag{3.23}$$

illustrate the typical energy needs and usage patterns of the community.

The average daily peak

$$\overline{P}_{d,+} = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(\max_{t \in d} E_B(t) \right), \tag{3.24}$$

and the average daily valley

$$\overline{P}_{d,-} = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(\min_{t \in d} E_B(t) \right), \tag{3.25}$$

provide insight into daily power usage swings.

The absolute maximum peak

$$P_+ = \max_{t=0}^{T} E_C(t), \tag{3.26}$$

and the absolute minimum valley

$$P_- = \min_{t=0}^{T} E_C(t), \tag{3.27}$$

provide information on the necessary line capacity and peak swing.

The average daily ramping rate

$$\overline{R}_d = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(\sum_{t \in d} |\nabla E_B(t)| \right), \tag{3.28}$$

provides a measure of momentary volatility of the net load signal of the community.

The load factor L indicates the efficiency of energy consumption with respect to peak load, ranging between 0 (inefficient) and 1 (most efficient), over a given period of time. Similar to Nweye et al. [71], a load factor complement $(1 - L)$ is reported in this section so that lower magnitudes are desirable across all metrics. Specifically, for the period of a day

$$1 - L_d = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(1 - \frac{\text{mean}_{t \in d} E_B(t)}{\max_{t \in d} E_B(t)} \right), \tag{3.29}$$

and for the period of a month

$$1 - L_m = \frac{1}{n_m} \sum_{m=0}^{n_m} \left(1 - \frac{\text{mean}_{t \in m} E_B(t)}{\max_{t \in m} E_B(t)} \right), \tag{3.30}$$

where n_m is the number of months, and m denotes a specific month.

Experiments are conducted to compare the performance of ALEX as a DER management system (DERMS) on the CityLearn2022 data set with a set of baselines. This study primarily focuses on testing the proposed hypotheses, deferring the investigation into the potential of ALEX as a state-of-the-art DERMS to future work.

ALEX utilizes building-level data to optimize a single, building-level objective. Consequently, we avoid benchmarking against algorithms that utilize community-level information or employ multi-objective optimization. To evaluate the hypotheses stated in Section 3.3.1, ALEX is compared against two benchmarks

- **NoDERMS**: The standard CityLearn2022 community, where no building exploits its battery storage capacities, serves as the performance baseline for our experiments.

- **IndividualDERMS**: In this case, a 'smart' net billing scenario is considered for the CityLearn2022 community, where each building maximizes self-sufficiency, prioritizing the reduction of building-level peaks and valleys while minimizing the ramping rate. The building policies for IndividualDERMS are generated using the same approach as the building policies for ALEX, i.e., by modifying the reward function to incentivize self-sufficiency with minimal building peaks and valleys.

The NoDERMS scenario is included in the experimental results and discussion to allow comparison with studies reporting normalized scores of metrics on the CityLearn2022 data set. Both scenarios together form reasonable benchmarks to evaluate our hypotheses. For Hypothesis 1 to be valid, ALEX should strictly improve all metrics compared to the NoDERMS scenario. Assuming Hypothesis 2 holds, ALEX is expected to surpass the IndividualDERMS in terms of average daily imported energy $\overline{E}_{d,+}$ and average daily exported energy $\overline{E}_{d,-}$. This is achievable only through a more effective use of the community's load-shifting capability, redistributing surplus energy from one building to another with spare battery capacity. For ALEX to outperform IndividualDERMS for all established metrics, both formulated hypotheses must be true.

3.4 Results and Discussion

The community's average daily net load profile is shown in Figure 3.1.

As anticipated, the IndividualDERMS noticeably flattens the average daily net load of the community compared to the NoDERMS scenario. ALEX, in turn, exhibits further improvement in this aspect, demonstrating a reduced swing with a significantly diminished valley.

Table 3.1 shows the results for each scenario in terms of metrics, facilitating their quantitative analysis.

Metric	NoDERMS	IndividualDERMS	ALEX
$\overline{E_{d,+}}$	258.54	214.81	**202.68**
$\overline{E_{d,-}}$	-77.48	-26.49	**-12.46**
$\overline{P_{d,+}}$	25.61	19.95	**19.44**
$\overline{P_{d,-}}$	-16.55	-6.35	**-1.67**
P_+	49.06	**42.37**	**42.37**
P_-	-37.86	-36.80	**-29.34**
$\overline{R_d}$	4.28	2.87	**2.84**
$1 - L_d$	0.73	0.65	**0.64**
$1 - L_m$	0.82	0.80	**0.78**

Table 3.1: Summarized metrics for full simulation on CityLearn2022 data set [70] for NoDERMS, IndividualDERMS and ALEX scenarios. For description of the metrics c.f. Section 3.3.4. Best values are typeset in **bold**.

The performance values of ALEX, compared to the NoDERMS scenario, clearly support the validity of Hypothesis 1. The consistent improvement across all metrics, driven by participants' selfish bill minimization, indicates a strong alignment between

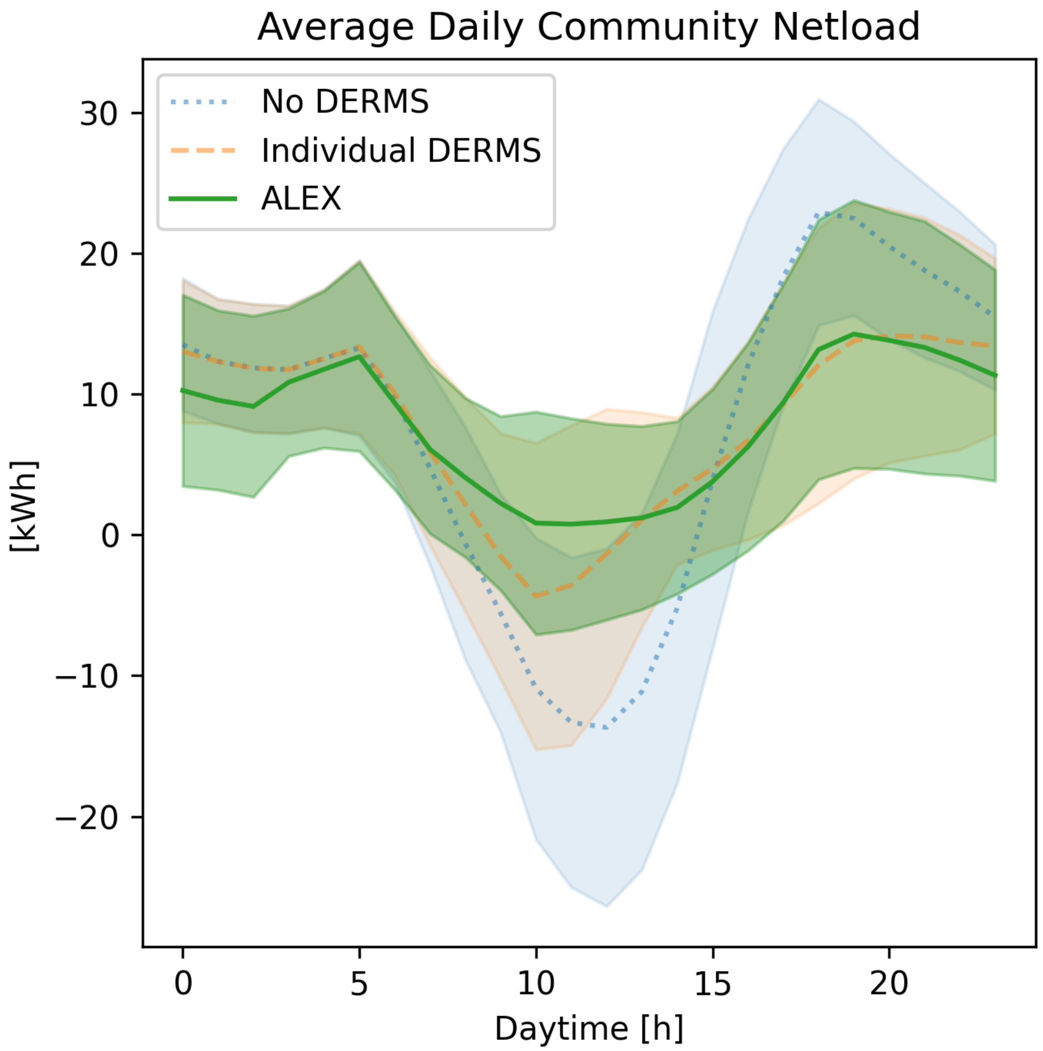

Figure 3.1: Average daily community net loads in kWh at hourly resolution for a full simulation on CityLearn2022 data set [70] for NoDERMS, IndividualDERMS, and ALEX scenarios. The plot displays both the average values and the standard deviation bands.

participant and grid stakeholder interests.

ALEX significantly reduces average daily exports and imports. In comparison to IndividualDERMS, ALEX consumes a higher proportion of locally generated energy. It is essential to note that IndividualDERMS optimizes for building-level self-consumption. Therefore, ALEX's improvement in average daily imports and exports can be solely attributed to its capacity to utilize the unused shifting capabilities of the community when some buildings have spare battery capacity, and others have surplus generation. This strongly supports Hypothesis 2. This observation is further supported by the graphs of the average daily *SoC* profiles in Figure 3.2.

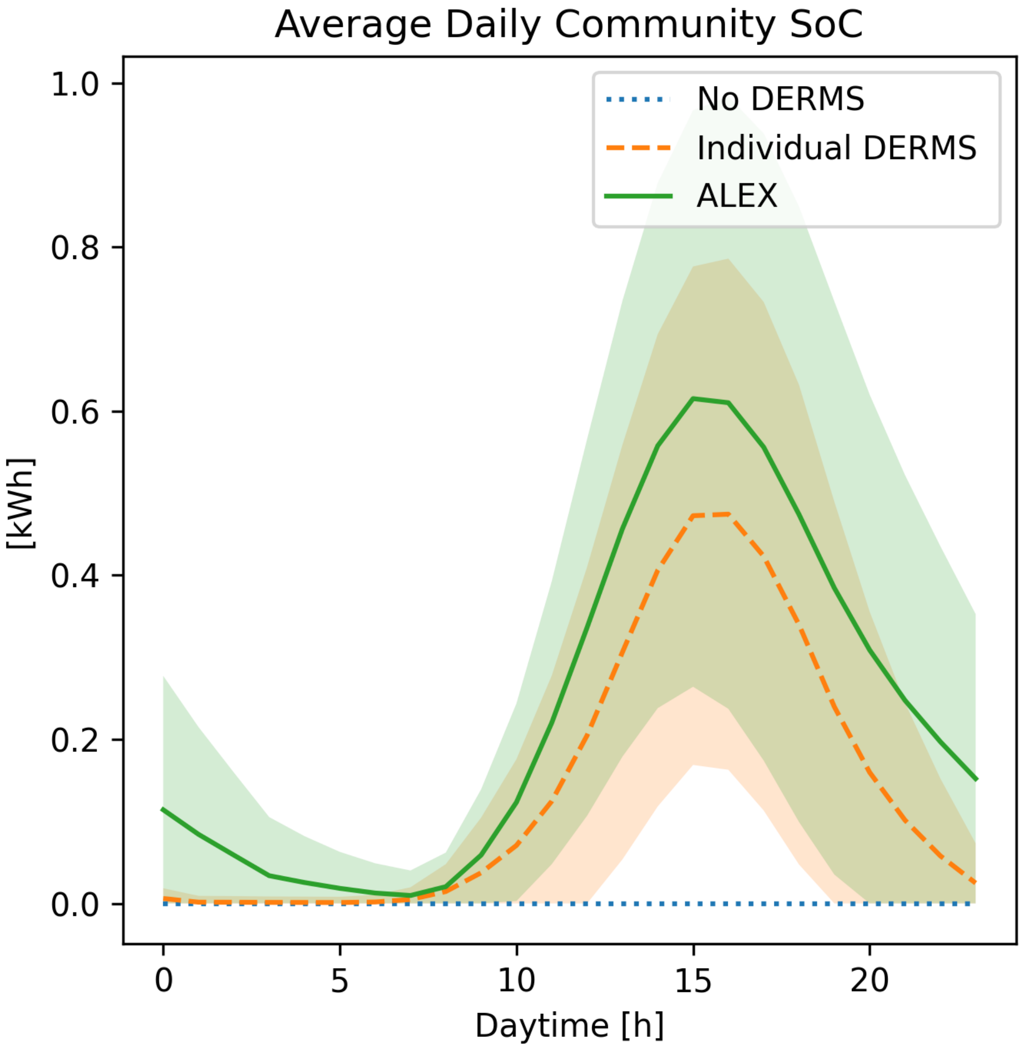

Figure 3.2: Average daily community *SoC* values are presented at hourly resolutions for a full simulation on the CityLearn 2022 dataset [70], encompassing NoDERMS, IndividualDERMS, and ALEX scenarios. The figure displays both the average values and the standard deviation bands.

The increased swing in *SoC* for ALEX corresponds to greater utilization of the community's load-shifting capacity. In contrast, IndividualDERMS is designed to maximize battery utilization at the building level. Therefore, the heightened battery utilization in ALEX must stem from community-level DER resource coordination, i.e., coordination between buildings within the community. Figures 3.1 and 3.2 confirm that ALEX equalizes the load in the community in a constructive manner. By summing the average daily import $\overline{E}_{d,+}$ and the average daily export $\overline{E}_{d,-}$, the average total energy consumed by the community per day can be calculated. The NoDERMS scenario community consumes 181.06 kWh, which is less than the IndividualDERMS community with 188.32 kWh. The ALEX community consumes 190.22 kWh, which is more than the IndividualDERMS community. Given that the battery energy storage system has an efficiency less than 100%, any energy temporally shifted within the community to satisfy later demand must compensate for incurred round-trip and self-discharge losses. Consequently, the community utilizing its shifting capabilities the most will also exhibit the highest net energy consumption, along with the lowest average daily exported and imported energy. This finding further confirms Hypothesis 2.

The evaluation of Hypothesis 2 is conducted by assessing average daily peak, maximum peak, maximum valley, community ramping rate, and load factor complements. ALEX outperforms IndividualDERMS for all metrics except the maximum peak, where both ALEX and IndividualDERMS perform equally. Given that IndividualDERMS minimizes peaks and valleys at the building level, the reductions in average daily peak and valley by ALEX are particularly significant. Unlike the NoDERMS scenario, IndividualDERMS and ALEX reduce the average daily peak $\overline{P}_{d,+}$ by 22.1% and 24.1%, respectively, and the average daily valley $\overline{P}_{d,-}$ by 61.6% and 89.9%, respectively. Although both alternatives equally reduce the maximum peak P_{+}, ALEX significantly reduces the minimum valley P_{-} by 22.5%, compared to IndividualDERMS, which achieves only a 2.7% reduction. This confirms that ALEX is more efficient at load balancing across the community. Similarly, in terms of community net load volatility, ALEX consistently outperforms IndividualDERMS, achieving higher reductions in ramping rate $\overline{R_d}$, daily load factor complement $1 - L_d$, and monthly load factor complement $1 - L_m$. This leads to the conclusion that the ALEX-managed system continuously maintains a better-behaved, less variable community net-load curve than the benchmarks, resulting in a more stable local electricity grid.

The performance values and quantitative analysis strongly support both hypotheses. These experiments robustly demonstrate that ALEX, assuming rational actors, exhibits all the desirable features of a LEM: ALEX aligns electricity end-user interests with grid stakeholder interests, as the act of maximizing relative profits (minimizing bills and maximizing DER-related returns) strongly correlates with improvements in various metrics indicative of electricity system stability. The community-level coordination to achieve such effects is present, despite each automating agent having access only to building-level information. This allows ALEX to exhibit properties usually associated with centralized DR approaches. In essence, ALEX as a LEM provides a pathway to implement TE at the grid-edge.

3.5 Summary and Conclusion

This study investigates ALEX, a TE-based LEM, where rational agents automate building DER management and trading. Each agent represents one building and aims to minimize its electricity bill using only building-level information. The ALEX-specific LEM mechanism is purely economy-driven and encourages rational agents to price in relation to the current round's supply/demand ratio. The concept of LEM as a tool for implementing community-level TE has gained traction recently. The successful implementation of such systems would address a growing, emerging set of DER-related challenges that grid stakeholders face.

Despite its promise, recent literature has shown that for LEMs, the common objective of maximizing relative profits might not necessarily result in the originally intended DR behavior but instead produce adverse effects [68, 69]. This could be a result of a market mechanism that insufficiently aligns participant and grid-stakeholder interests or insufficiently tuned participant heuristics, for example. A common strategy to ensure alignment between participants and grid stakeholders is explicitly considering grid performance or related metrics in the LEM's price formation process. Taking this information into account, this study aims to investigate the following two hypotheses:

Hypothesis 1: ALEX's market mechanism, which incentivizes bid and ask prices in correlation to the current timestep's supply and demand ratio within a profitability gap, is capable of fostering a strong alignment between participant and grid stakeholder interests;

Hypothesis 2: Rational agents, representing individual buildings within ALEX, coordinate DER usage patterns across the entire community B, despite each agent operating with only building-level information and selfishly minimizing its own electricity bill.

Both hypotheses are tested through a set of experiments designed to benchmark ALEX with fully rational agents against a baseline NoDERMS approach and an IndividualDERMS approach that maximizes self-consumption, while minimizing the ramping rate and the peak net load at the building level. This comparison is performed using the CityLearn2022 dataset, and the performance of both approaches is assessed using a suite of community net load metrics indicative of the state of the local electricity grid, such as ramping rate, load factor, peak export and import, and average daily export and import. The behavior of ALEX rational agents is simulated with an algorithm that combines iterative best response with dynamic programming through value iteration.

The experimental results confirm both hypotheses. ALEX's settlement mechanism appears sufficient to generate alignment between participants' selfish financial interests and grid stakeholders' interests, thereby improving the local performance of the electricity grid despite being economy-driven. All load-balancing and smoothing properties result from bill minimization, as agents are neither explicitly incentivized to coordinate nor optimize for any of the investigated metrics. ALEX exhibits community-level coordination of DERs and outperforms the IndividualDERMS baseline across all investigated metrics. These experiments demonstrate that ALEX is a decentralized DERMS with properties usually associated with centralized DR approaches, such as community-level coordination.

In addition to demonstrating that economy-driven LEM such as ALEX have the potential to successfully deliver on the promise of LEM, this article contributes to closing several research gaps in the current LEM literature. The simulation approach for rational actors developed in this article can be applied to other LEM designs, addressing the unreliability of LEM investigation using expert-designed agent heuristics. The CityLearn2022 dataset is a high-quality, benchmarkable, open-source dataset that has been previously applied to non-LEM DERMS. Its successful application in the LEM environment described in this article is an additional contribution toward establishing an accepted benchmark dataset for DERMS.

The main focus of future work is to enhance the research by training a group of generalizing rational actors using state-of-the-art DRL techniques and evaluating their effectiveness as DERMS. This exploration will open up several areas for additional investigation. For example, it provides an opportunity to examine the differences between single-agent and multi-agent setups, investigating how various configurations impact dynamics and performance. Moreover, the investigation will include the exploration of different methods for generating and handling rewards, providing valuable insights into the developing field of decentralized energy management.

Chapter 4

Decentralized Coordination of Distributed Energy Resources facilitated by Local Energy Markets and Deep Reinforcement Learning

4.1 Introduction

Progress towards sustainable energy utilization is crucial for addressing climate change. In this context, the convergence of technological advances and lagging regulatory frameworks has precipitated the rapid adoption of distributed energy resources (DERs), reshaping the dynamics of the grid edge where electricity end-users reside [4]. Consequently, the variability of the net load at the grid edge is rapidly increasing. The term variability encompasses the composite effects of intermittency and other net load volatilities, such as those caused by electric vehicle charging. This marked increase amplifies the challenges associated with ensuring the reliability and efficiency of grid operations [79, 80]. This drives the transition to the Smart Grid, which operates in a decentralized and autonomous manner to maintain and possibly enhance the operability of the electricity grid.

To address these challenges, the research community has been actively exploring demand response (DR) methodologies. Broadly speaking, DR techniques leverage various signals to modulate end-user load demand, supporting electrical grid efficiency and reliability. These signals encompass both direct control commands to assets and incentive mechanisms intended to influence end-user behavior, thus delineating between direct and indirect DR. Notably, the key hurdles in indirect DR lie in aligning the interests of grid stakeholders and electricity end-users through appropriate incentive structures and subsequently ensuring sufficient participation to achieve the desired effect [8, 9, 21].

Traditionally, schedule-based approaches employing model predictive control (MPC) frameworks have been predominant in indirect DR. These approaches rely on behavioral models to form a forecast and then attempt to optimize load demand over a future time horizon. However, their inherent reliance on expert knowledge, high time complexity, and bias toward centralized information processing may impede their efficacy in addressing the rapid and disparate changes observed at the grid edge.

In response to the challenges faced by these scheduling-based methods, transactive energy (TE) has emerged as a compelling alternative. TE, defined as "the use of a combination of economic and control techniques to improve grid reliability and efficiency" by the GridWise Architecture Council [10], aligns well with the Smart Grid ethos, emphasizing the market as a decentralized delivery mechanism for incentive signals [9].

Recent literature has highlighted the concept of Local Energy Markets (LEMs) as a viable path to implement TE within geographically constrained communities at the grid edge. Mengelkamp et al. define LEM as "a geographically distinct and socially close community of residential prosumers and consumers who can trade locally produced electricity within their community. For this, all actors must have access to a local market platform on which (buy) bids and (ask) offers for local electricity are matched" [20]. LEMs allow for the delivery of real-time incentive signals to electricity end-users, providing the necessary granularity and immediacy within a decentralizable framework.

The surveys of completed DR pilot studies confirm that automation is necessary to facilitate sufficient levels of participation [8, 9, 21]. While MPC is entrenched in the general DR literature for automation, model-free approaches such as deep reinforcement learning (DRL) present a promising paradigm better suited to tackle the challenges faced at the grid edge. Initially inspired by high-level performance showcases of DRL in games [81–83], this notion is reinforced by the success of DRL in fields like robotics [24] and process control [25]. Moreover, it is supported by a growing body of research applying DRL to the electricity grid [27–29].

Within this context, recent studies have explored automating end-user participation and DER management in LEMs [13–15, 17, 18, 49, 84], predominantly through agents trained to optimize end-user bills via load-shifting capacities. Some studies demonstrate the reduction of net community energy consumption [18, 84], while others investigate the provision of flexibility services [14]. However, to the best of the authors' knowledge, there are no other studies demonstrating the reduction of community-level load variability through the automation of LEMs using DRL.

Such a conclusive demonstration is not trivial. Despite the intention of LEMs to align the interests of end-users with the objectives of grid stakeholders, it is crucial to recognize that incentivized behavior may not automatically translate into reduced variability or enhanced power quality at the local level [68, 69]. Similarly, the intricate interplay between LEM design and participant automation may yield unforeseen outcomes [47], a phenomenon commonly observed when automating complex systems using DRL [30].

This article addresses this research gap by training independent agents to automate end-user participation in LEMs and the utilization of DERs. The study demonstrates an emergent reduction in community net load variability even when agents solely prioritize individual bill optimization. To enhance benchmarking and future comparability, performance evaluation is conducted on an open-source dataset, and the agent's performance is compared to several baselines. The trained DRL agents perform close to the near-optimal benchmark without information sharing or access to future information.

Subsequent sections of this article delve into related work and background in Section 4.2, methodology for training DRL agents, evaluation and benchmarking procedures in Section 4.3, a comprehensive discussion of simulation results in Section 4.4, and conclude with a brief summary and avenues for future research in Section 4.5.

4.2 Related Work and Background

Subsection 4.2.1 briefly reviews related literature and establishes a notable research gap: the lack of a well-benchmarked demonstration of variability reductions within an economy-driven LEM, emerging from selfish end-user bill minimization that DRL agents automate. Subsection 4.2.2 introduces the LEM design that forms the foundation of this study. Subsection 4.2.3 overviews reinforcement learning and proximal policy optimization, the base DRL algorithm employed within this article.

4.2.1 Related Literature

The application of DRL in DR, and for the electricity grid in general, has garnered significant attention in recent years [27–29]. Studies exploring the distributed coordination of DERs through DR mechanisms outside of LEM, such as those by Chung et al. [11], Zhang et al. [12], and Nweye et al. [71], tend to optimize for composite rewards and incorporate community-level metrics related to grid stability or variability, following a direct optimization approach.

Concurrently, there has been a surge in literature investigating LEMs. Mengelkamp et al. [20], Capper et al. [19], and Tushar et al. [21] provide comprehensive insights into the evolving LEM ecosystem. In general, this field tends to focus on the socioeconomic performance of the proposed system, while DR aspects are only narrowly discussed, and performance benchmarking tends to be restricted.

For instance, Liu et al. [85] propose a LEM-like mechanism, using pricing based on the supply-demand ratio to coordinate energy flow between microgrids, leveraging MPC for automation. Similarly, Lezama et al. [86] explore LEMs from a grid integration perspective, focusing on socioeconomic performance. Ghorani et al. [87] develop bidding models for risk-neutral and risk-averse LEM agents, evaluating their socioeconomic efficacy under various market designs. Meanwhile, Mengelkamp et al. [47] investigate different market designs using heuristic agents, focusing on socioeconomic metrics. A burgeoning body of research emphasizes the automation of LEM participation through DRL. Xu et al. [18] employ a MARL Q-learning algorithm to automate participation in a LEM that communicates a pricing schedule based on a supply and demand forecast. Zhou et al. [17] propose an economy-driven LEM pricing mechanism, optimizing participant bidding via a combination of Q-learning and fuzzy logic. Similarly, Zang et al. [49] train end-user agents to interact with community-level batteries within LEMs.

As Mengelkamp et al. [47] highlight, the integration of LEMs and automated participation presents complex challenges and potentially unforeseen consequences due to the emergent, intricate system dynamics. Investigations by Kiedanski et al. [68] and Papadaskalopoulose et al. [69] demonstrate that increases in socioeconomic performance in such settings may not directly translate to improved grid performance in terms of reducing variability or improving power quality.

To address this issue, some studies incorporate electricity grid performance metrics into the LEM's pricing mechanism or the agent's reward function, diverging from the original purely economic focus of LEMs and adopting a direct optimization approach. For example, Chen et al. [84] investigate microgrid trading in the context of LEMs, employing a reward function with explicit constraints. Their findings demonstrate that this approach increases self-sufficiency compared to expert-designed heuristics and random action agents in benchmarking experiments. Similarly, Ye et al. [14] explore the use of LEMs to provide flexibility services. Their contribution stands out by benchmarking against a near-optimal MPC baseline, establishing a reasonable upper performance limit. However, even such contributions do not evaluate their agents' performance on variability-related metrics for which the agents do not explicitly optimize.

The principal promise of LEM, and, in a more general sense, TE, lies in the notion that a well-designed market mechanism should incentivize a broad range of beneficial behaviors. The underlying ambition is to achieve this without explicitly tying the market's cost function to these outcomes, enabling agile and robust decentralization by avoiding the need for expensive real-time computation of an expressive set of related metrics. In a sense, optimizing end-user bills should indirectly and emergently reduce net load variability in this setting. Despite the current landscape of contributions, the demonstration of such behavior via an LEM that relies on DRL for automation purposes is still outstanding. This study aims to contribute to closing this research gap.

4.2.2 Autonomous Local Energy eXchange

The Autonomous Local Energy eXchange (ALEX), initially proposed by Zhang et al. [1], serves as an LEM for a community denoted as B, where individual buildings $b \in B$ participate in energy trading facilitated by a round-based, futures-blind double auction settlement mechanism. In the context of a round-based futures market, trading occurs in predefined time intervals. A futures market accepts bids and asks for a future settlement timestep t_{settle}, to be submitted at the current time step t_{now}. Settlements are then executed at a subsequent time step t_{deliver}, with $t_{\text{deliver}} > t_{\text{settle}} > t_{\text{now}}$. In such a blind double auction market, each building interacts with the market without awareness of other buildings' activities.

In ALEX, market participants do not share information and instead selfishly optimize their individual electricity bills. The building's electricity bill consists of two main components: the market bill and the grid bill. The market bill includes the cumulative settlement cost, determined by pairing each settlement price with its corresponding quantity for the building. In contrast, the grid bill covers the residual amount required to meet the household's energy demand, billed at the prevailing grid rate selling or buying price, depending on the current net-billing scenario. A profitability margin between the grid rate selling and buying prices serves as an incentive for LEM utilization. This means that any exchange over the LEM presents a favorable scenario. Achieving this could involve leveraging tracked greenhouse gas emission savings or partial fee offsets [19, 67].

Zhang et al. [1] delve into the essential properties required for ALEX's settlement mechanism to incentivize RL agents to learn pricing in correlation with the settlement timestep t_{settle} supply and demand ratios. Formulating ALEX as a mixed-form stochastic game suggests the existence of at least one Nash equilibrium. This insight facilitates the identification of a market mechanism possessing the desired properties through experiments that employ tabular Q-learning bandits under varied but fixed supply and demand ratios. Subsequent experiments deduce a market price function based on the current supply and demand ratio. A follow-up study by Zhang and Musilek [75] investigates a system incorporating a communal battery energy storage system (BESS) controlled by an expert-designed heuristic. The study demonstrates efficacy in avoiding violations of voltage-frequency constraints on a test circuit.

May and Musilek [2] further examines ALEX as a DR system. The authors simulate a group of near-optimal, rational actors on ALEX using an iterative best-response and dynamic programming algorithm. Their performance is then compared against several baselines. The identified policies reveal emergent community-level coordination of DERs, driven by incentives within the LEM. Remarkably, this coordination occurs even though each participant accesses only building-level information and selfishly optimizes their electricity bills. Consequently, these policies consistently outperform the benchmark building-level DR system across various community net-load-related metrics measuring net load variability at the community level. While this agent behavior shows promise, it is important to note that it is generated using a pure search approach that relies on a perfect forecast of end-user generation and demand.

This study aims to extend these results by training a set of DRL agents on the equivalent task without access to perfect forecasts, yet achievening a comparable level of variability reduction. This would effectively address the research gap identified in subsection 4.2.1.

4.2.3 Reinforcement Learning

Reinforcement Learning (RL) is a machine learning framework closely linked to optimal control paradigms.

As illustrated in Figure 4.1, RL focuses on optimizing the behavior of an agent that interacts with the environment through actions and subsequently receives observations and rewards.

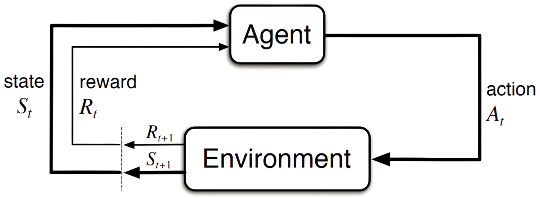

Figure 4.1: Agent to environment interaction diagram, taken from Sutton & Barto [48].

This is typically formalized through the markov decision process (MDP), represented by the tuple (S, A, P_a, R_a). The MDP encapsulates the state space S, action space A, transition probabilities P_a from state s to the next state s' upon taking an action a, and receiving an immediate rewards R_a. A policy, denoted as π, characterizes an agent's behavior through a probabilistic mapping from state s to action a. For instance, this mapping could take the form of a Gaussian distribution, where the mean μ and standard deviation σ are functions of the state s.

MDPs within the context of RL are typically time-discrete, allowing the notation of the time-step t to represent a specific point in the interaction trajectory between the agent and the environment. This trajectory starts at $t = 0$ and concludes at $t = T$. The return G signifies the cumulative, discounted future reward,

$$G_t = \sum_{t=0}^{T} \gamma^t R_{t+1}, \tag{4.1}$$

which facilitates the definition of state value

$$V_\pi(s_t) = \mathbf{E} G_t \forall \pi, \tag{4.2}$$

and state-action value

$$Q_\pi(s_t, a_t) = \mathbf{E} G_t \forall \pi, \tag{4.3}$$

where γ is the discount factor. The primary objective is to identify an optimal policy π^* which maximizes the expected return $\mathbf{E}G$.

Distinguished from other MDP search methods by its emphasis on temporal difference and bootstrapping, RL agents iteratively learn the optimal policy π^*. They adjust their encoding in response to the reward signal received from the environment. The parameters underlying this encoding are denoted as θ and are updated through an RL learning algorithm's loss function, often employing a stochastic gradient descent method. RL algorithms are generally categorized into two types: value-based and policy gradient methods. Value-based methods estimate state values V or state-action values Q and subsequently associate policies π with these estimates. On the other hand, policy gradient methods directly learn policies π or their parameters using a policy loss

$$L(\theta) = \mathbb{E}\left[\log \pi_\theta(a_t, s_t) V_t\right], \tag{4.4}$$

with actor-critic methods utilizing a critic to estimate state values V and compute advantages A in order to reduce variance, resulting in the corresponding actor-critic loss

$$L(\theta) = \mathbb{E}\left[\log \pi_\theta(a_t, s_t) A_t\right], A_t = V_t - V_\theta(s_t). \tag{4.5}$$

Deep Reinforcement Learning (DRL), an amalgamation of RL and deep neural networks, has gained traction for its ability to solve complex MDPs in a generalized manner [81–83]. DRL methods leverage replay buffers to store agent-environment interactions, facilitating multiple mini-batch stochastic gradient descent epochs. This necessitates the differentiation between the parameter set used to collect samples into the replay buffer θ_{old} and the new parameters θ, which emerge as a result of gradient updates.

Particularly noteworthy within the dynamic landscape of DRL is Proximal Policy Optimization (PPO), introduced by Schulman et al. [31]. In contrast to naive actor-critic approaches, PPO employs a clipped surrogate objective based on the probability ratio $r(\theta)$. This ratio compares the probabilities of the new policy π_θ and the old policy $\pi_{\theta_{old}}$, aiming to mitigate policy drift and ensure the reliability of data collected into the replay buffer. PPO's actor loss clips the magnitude of the policy ratio $r(\theta)$ within a tolerance parameter ϵ

$$L(\theta) = \mathbb{E}\left[\min\left(r(\theta)A_t, clip(r(\theta), 1-\epsilon, 1+\epsilon)A_t\right)\right]. \tag{4.6}$$

In addition, most Proximal Policy Optimization (PPO) implementations incorporate generalized advantage estimation, a technique proposed by Schulman et al. [88], to reduce the variance of the advantage A.

4.3 Methodology and Evaluation

This study aims to extend previous contributions [1, 2] by training DRL agents to autonomously participate in ALEX. The expectation is that these agents will demonstrate a level of emergent community-level variability reduction that is comparable to the near-optimal search method described by May and Musilek [2], but without relying on a perfect forecast. Such a showcase of variability reduction within a DRL-driven LEM context would address the significant research gap outlined in Subsection 4.2.1.

To achieve this goal, this section formulates ALEX environment as MDP in Subsection 4.3.1, outlines the DRL algorithm employed for training the agents in Subsection 4.3.2, and elucidates the experimental design in Subsection 4.3.3. The latter also includes details on evaluation performance metrics and baselines.

4.3.1 Autonomous Local Energy eXchange as Markov Decision Process

The formulation of ALEX as an MDP involves defining the agent's observations O, actions a, rewards r, and policy π. In comparison to the initial formulation [2], the approach outlined here incorporates specific adaptations tailored to the nature of ALEX as a futures market. This is crucial, given the constraint that the DRL agents should not rely on future information. Additionally, the formulation accommodates continuous observation and action spaces for the DRL agents.

The individual agent's MDP encapsulates the viewpoint of a single agent within the ALEX environment. Given that participants in ALEX neither share information nor engage in communication, this individual agent MDP is partially observable. This contrasts with the fully deterministic nature of the joint MDP. In this study, the DRL agents must function as fully independent actors, navigating a continuous action and a partially observable, continuous state space. Accordingly, this section adopts this perspective and refers to the state space S as the observation space O.

The observation space O^b for an individual agent at timestep t encompasses various continuous variables, including the current net load E_t^b, battery state of charge SoC_t^b, the average last settlement price $p^{bid}_{t_{\text{last settled}}}$, and total bid and ask quantities from the last settlement round $q^{bid}_{t \text{last settled}}$ and $q^{ask}_{t_{\text{last settled}}}$, respectively. To capture temporal patterns such as daily and yearly seasonalities, sine and cosine transformations of the current timestep t are incorporated instead of using the raw timestamp.

$$\begin{aligned} O_t^b := (\ & sin(t)_{year}, cos(t)_{year}, sin(t)_{day}, cos(t)_{day}, \\ & E_t^b, SoC_t^b, p_{t_{\text{last settled}}}, q^{bid}_{t_{\text{last settled}}}, q^{ask}_{t_{\text{last settled}}}\ ) . \end{aligned} \tag{4.7}$$

However, future information, such as net load at settlement time $E^b_{t_{\text{settle}}}$, is not included in this observation space.

In contrast to the action space proposed by Zhang et al. [1], the action space A^b for an agent at timestep t exclusively includes the continuous battery action, scheduled for the future settlement time step $a_{BESS,t_{\text{settle}}}$. This action is constrained by the battery's charge and discharge rates. The determination of bid and ask quantities at settlement time t_{settle} relies on the residual net load, while bid and ask market conditions dictate prices following the round's closure, guided by the price curve defined by Zhang et al. [1].

The building's battery action $a^b_{BESS,t\text{settle}}$ is defined as a superposition of two components: the self-sufficiency maximizing, greedy battery action $a^{\pi_0}_{BESS,t_{\text{settle}}}$ and the agent's learned action $a^{\pi_\theta}_{BESS,t_{\text{settle}}}$. Here, the policy π_0 represents the self-sufficiency maximizing policy, which aims to greedily minimize the amplitude of the participant's net load E^b_t using the residential BESS.

$$a^b_{BESS,t_{\text{settle}}} := a^{\pi_0}_{BESS,t_{\text{settle}}} + a^{\pi_\theta}_{BESS,t_{\text{settle}}}. \tag{4.8}$$

This action and agent policy definition offers several distinct advantages, significantly expediting the learning process of the studied DRL agents. The policy π_0 can be computed at settlement time and serves as a reasonable initial heuristic, even though it may be far from the optimal policy. This approach enables more efficient state exploration while mitigating some of the internal environment modeling that the agent has to perform.

As a result, the agent's reward function is formulated as the difference between the electricity bill $bill^b_t$ and the bill incurred by the self-sufficiency maximizing policy π_0, denoted as $bill^{b,\pi_0}_{t\text{settle}}$. This approach, in contrast to using the naive participant electricity bill $bill^b_t$ as a reward signal, offers a clearer indication of whether the RL agents are learning a useful policy

$$r^b_t := bill^b_{t_{\text{settle}}} - bill^{b,\pi_0}_{t_{\text{settle}}} \tag{4.9}$$

4.3.2 Shared Experience Recurrent Proximal Policy Optimization

The agents in this study undergo training as independent agents with shared experience [89]. Although each agent acts autonomously and solely accesses building-level information, they aggregate trajectories into a shared replay buffer. During trajectory collection, the actors function as independent copies of the same actor and critic neural network, which is updated from the shared replay buffer. This maintains full independence between agents during rollout but promotes faster convergence. Christianos et al. [89] demonstrated the efficacy of this approach in enhancing performance within complex multi-agent environments when compared to a fully independent learning setup. Observations undergo standardization and mean-shifting, while rewards are solely standardized, following best practices proposed by Schulman et al. [90].

The remaining portion of this section details modifications to the underlying PPO algorithm. A recurrent PPO [91], using a Long Short-Term Memory (LSTM) [92] hidden layer for both the actor and the critic, is enhanced with recurrent burn-in and initialization, proposed by Kapturowski et al.[32]. Drawing motivation from the findings of Andrychowicz et al. [34], after processing a replay buffer, the new weights θ are used to recalculate the hidden states of the agent LSTM based on the entire trajectory experienced during the current episode. Both enhancements address the risk of stale or drifted state representations, enhancing the agent's capacity to develop meaningful state representations and a long-term context. Informed by Ilyas et al. [33] and with the goal of convergence towards a Nash equilibrium, the learning rate is annealed throughout the training. Instead of setting the value for the terminal transition at T to 0, this study takes it from the critic's value prediction, with preliminary tests indicating an accelerated convergence of the critic to a higher explained variance. Furthermore, instead of naively imposing action space boundaries by clipping the Gaussian distribution, the algorithm used in this study employs a squashed Gaussian distribution followed by renormalization, as popularized by soft actor-critic algorithms [93].

The initialization of the actor's final layer is designed to ensure that the mean μ exhibits an expected value of 0. This is achieved by sampling the weights and biases of this layer from a uniform distribution between 0.001 and -0.001. In a similar vein, the policy's standard deviation σ is initialized very narrowly. This setup enables the agent to commence training based on trajectories collected near the self-sufficiency maximizing policy π_0. This strategy is grounded in the assumption that the optimal policy π_θ^* is much closer to the self-sufficiency policy π^0 than to a pure random policy. Large deviations from π^0 are considered highly situational, while smaller deviations are more common. From a task decomposition perspective, the RL agents learn how to load shift to maximize self-sufficiency, an internally focused task, and then proceed to learn how to leverage the market, an externally focused task. Hence, this practice aims to bias the agents to first learn how to load shift and then learn how to utilize the market. Both adjustments contribute to notable improvements in convergence for the studied task.

Hyperparameters are used in this study are provided in Appendix C.1, along with a brief discussion of the tuning and monitoring process.

4.3.3 Experimental Design

The DRL agents are trained and evaluated on the CityLearn2022 dataset [70]. The open-source nature of this dataset enables subsequent studies to directly benchmark against this contribution across a diverse range of DR applications. This dataset provides a year of hourly data for 17 smart community buildings, featuring time series of energy demand, photovoltaic generation, and BESS performance characteristics. For each building, one independently acting agent is trained as outlined in Subsection 4.3.2. Therefore, one episode is defined as a full trajectory over the dataset and lasts 8760 steps, while one run fully trains such a set of 17 agents. For evaluation, the parameter set θ with the best episodic communal return G^B is selected from a run, assuming that this snapshot represents the best-performing equilibrium between agents. This snapshot is updated throughout training when a new best communal return is achieved. From a set of 5 runs, the median performing run is selected for benchmarking purposes.

To assess agent performance, we employ a set of metrics from May and Musilek [2]. All performance metrics in this study are functions of the community net load E^B, defined as the summation of all building net loads E^b. The following expressions utilize n_d to denote the number of days in the dataset, d to represent the number of time steps in a day, and t as the current time step. The notations $\max_{\text{start}}^{\text{stop}}$ and $\min_{\text{start}}^{\text{stop}}$ denote the maximum and minimum operands over the interval from start to stop, respectively. Given the hourly resolution of the dataset used in this study, the conversion from kilowatt-hours (kWh) to kilowatts (kW) is excluded from the notation. The performance metrics encompass:

- The average daily imported energy

$$\overline{E}_{d,+} = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(\sum_{t \in d} \max(E^B(t), 0) \right) \tag{4.10}$$

- The average exported energy

$$\overline{E}_{d,-} = \frac{-1}{n_d} \sum_{d=0}^{n_d} \left(\sum_{t \in d} \min(E^B(t), 0) \right) \tag{4.11}$$

- The average daily peak

$$\overline{P}_{d,+} = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(\max_{t \in d} E^B(t) \right) \tag{4.12}$$

- The average daily valley

$$\overline{P}_{d,-} = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(\min_{t \in d} E^B(t) \right) \tag{4.13}$$

- The absolute maximum peak

$$P_+ = \max_{t=0}^{T} E^B(t) \tag{4.14}$$

- The absolute minimum valley

$$P_{-} = \min_{t=0}^{T} E^B(t) \tag{4.15}$$

- The average daily ramping rate

$$\overline{R}_d = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(\sum_{t \in d} |\nabla E^B(t)| \right) \tag{4.16}$$

- The daily load factor complement

$$1 - L_d = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(1 - \frac{\text{mean}_{t \in d} E^B(t)}{\max_{t \in d} E^B(t)} \right) \tag{4.17}$$

- The monthly load factor complement

$$1 - L_m = \frac{1}{n_m} \sum_{m=0}^{n_m} \left(1 - \frac{\text{mean}_{t \in m} E^B(t)}{\max_{t \in m} E^B(t)} \right). \tag{4.18}$$

This comprehensive set of metrics offers insights into the variance of the community net load E^B across various time scales. These time scales range from the hourly perspective, as captured by the ramping rate $\overline{R}_d$, to daily and monthly perspectives, as captured by the daily and monthly load factors $1-L_d$ and $1-L_m$. Additionally, the yearly and daily averages of peak load demands and generation values provide valuable information about community energy consumption and infrastructure strains. Importantly, all metrics are formulated so that lower values are preferable. Collectively, these metrics provide a robust framework for assessing the performance of an arbitrary DR system regarding its general impact on net load variability.

To effectively gauge the relative performance of the trained DRL agents, three benchmarks are used:

- **NoDERMS**: This baseline corresponds to the default community, where no building exploits its battery storage capacities. It serves as the reference setting and is expected to be easily outperformed by any DR system.

- **IndividualDERMS**: In this benchmark, each building in the community operates under a net billing strategy. Buildings prioritize self-sufficiency by smoothing building-level peaks and valleys while minimizing the ramping rate [2]. This benchmark serves as a reasonable performance baseline, resembling a well-tuned heuristic system commonly found in current DR applications. Importantly, unlike the proposed DRL agents, this benchmark has access to a perfect forecast.

- **ALEX DP**: This benchmark represents a near-optimal policy within a discretized version of ALEX's MDP. It is determined using a dynamic programming search method based on iterative best response and value iteration [2]. Importantly, unlike the proposed DRL agents, this benchmark has access to a perfect forecast.

The expectation is that ALEX RL shows a clear correlation between participant bill savings and improvement in the outlined performance metrics compared to the NoDERMS baseline. The desired outcome is for ALEX RL to perform comparably to ALEX DP. This achievement would indicate agent convergence to a near-optimal level of performance and a clear outperformance of the Individual DERMS benchmark. This outcome would effectively address the identified research gap by demonstrating a clear reduction in variability across the community due to participant automation DRL within a LEM. The achieved performance would be contextualized against a set of reasonable benchmarks.

4.4 Results and Discussion

This study aims to address a significant research gap highlighted in the background section by demonstrating a reduction in community-level variability of net load facilitated by DRL agents within a LEM. Towards this objective, this section establishes a clear connection between participant bill reduction and performance metrics within the chosen setting. Subsequently, a comparative analysis of the DRL agents against benchmarks introduced in the earlier subsection is conducted.

The training methodology of the agents focuses on their relative improvement compared to the self-sufficiency maximizing policy π^0, as outlined in Section 4.3.2. Convergence behaviors are visually depicted in Figure 4.2, highlighting the average building bill savings of ALEX RL across episodes, benchmarked against ALEX DP. The shaded area represents the variance between runs.

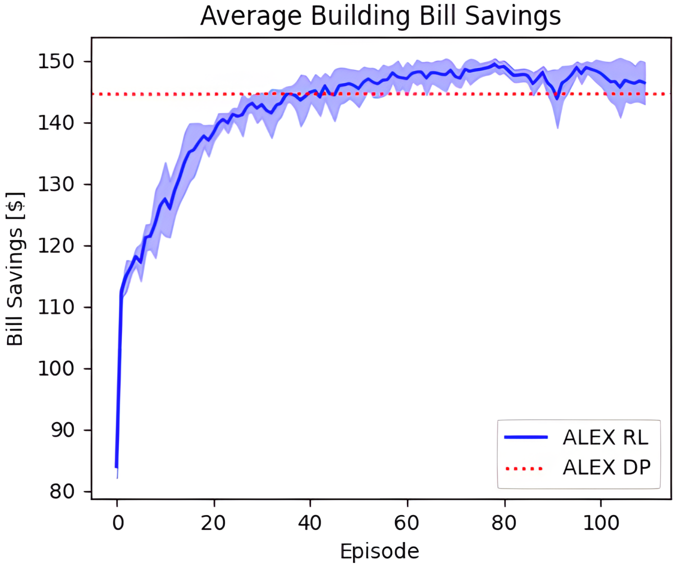

Figure 4.2: Average participant bill savings comparison between ALEX RL (blue), ALEX DP (red). Shaded areas depict variance bands between a set of 5 ALEX RL runs, trained over 117 episodes.

As evident from Figure 4.2, ALEX RL manages to achieve bill savings that slightly exceed those of ALEX DP. It is crucial to note that ALEX DP performs its search for one day ahead, while ALEX RL is not constrained in the duration of its load shifting. These results indicate that, for the CityLearn 2022 dataset, there is ample opportunity to shift load over several days.

To strengthen the correlation between achieved bill savings and evaluation metrics, Figure 4.3 tracks the performance of the median-performing run of ALEX RL in terms of performance evaluation metrics throughout training. A discernible downward trend is evident for all performance metrics, signifying a clear correlation between selfish bill minimization and the selected set of performance metrics.

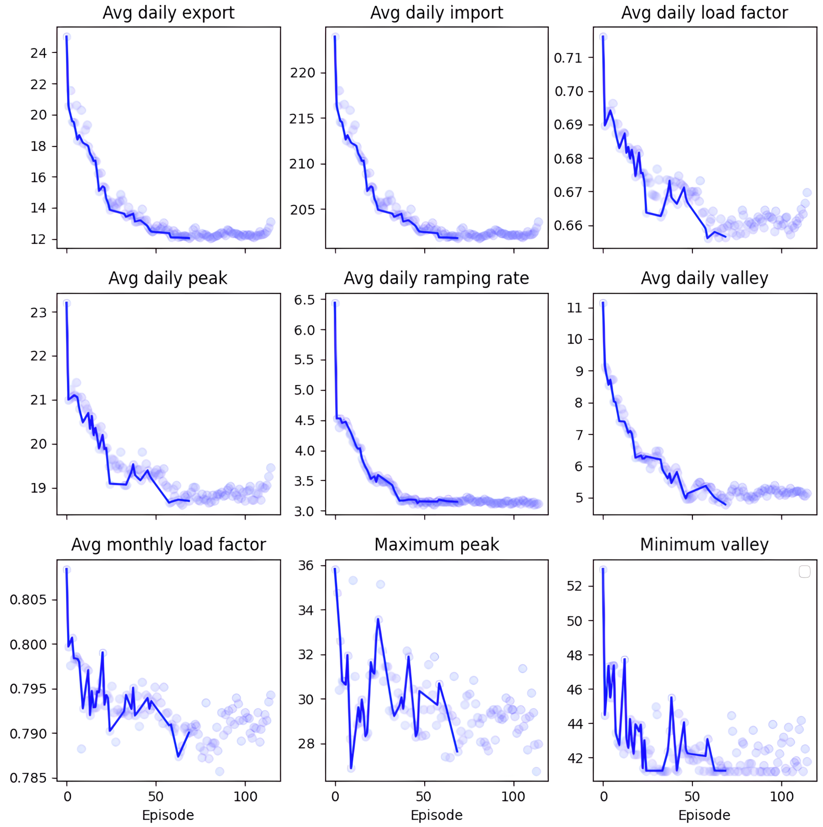

Figure 4.3: Performance of recorded community-level metrics per episode throughout training. The opaque scattered data points represent singular episode equivalents, while the blue line depicts the metric performance of the most recent highest return achieved.

Qunatitative analysis, summarized in Table 4.1, consistently supports correlations between performance metrics and participant bill savings. These findings affirm that training DRL agents within ALEX incentivize behavior conducive to the emergent suppression of variability in community net load.

These results strongly suggest that the observed correlations between performance metrics and return are consistent across runs. Furthermore, the observed maximum return correlations are consistently higher than the episodic equivalent. Considering

Metric Correlated to		Episodic Return	Maximum Return
Average daily import [kWh]	$\overline{E_{d,+}}$	-0.993 (-0.994)	**-0.994 (-0.995)**
Average daily export [kWh]	$\overline{E_{d,-}}$	-0.993 (-0.993)	**-0.994 (-0.994)**
Average daily peak [kW]	$\overline{P_{d,+}}$	-0.980 (-0.982)	**-0.982 (-0.982)**
Average daily valley [kW]	$\overline{P_{d,-}}$	-0.966 (-0.964)	**-0.975 (-0.972)**
Minimum peak [kW]	P_+	-0.466 (-0.470)	**-0.478 (-0.480)**
Maximum valley [kW]	P_-	-0.734 (-0.736)	**-0.775 (-0.770)**
Average daily ramping rate [kW]	$\overline{R_d}$	-0.934 (-0.932)	**-0.952 (-0.955)**
Average daily load factor	$1 - L_d$	-0.726 (-0.730)	**-0.833 (-0.833)**
Average monthly load factor	$1 - L_m$	-0.982 (-0.980)	**-0.985 (-0.982)**

Table 4.1: Pearson's correlations between the metrics and achieved bill savings; the rightmost column correlates Maximum Return episodes and their respective metric performance, while the middle column correlates episodic return and the respective episodic metric performance; the numbers in parentheses denote the average correlation over 5 training runs, whereas the non-bracketed number denotes the correlation of the run achieving the median return.

ALEX's nature as a mixed-form stochastic game, this outcome is not necessarily surprising and might result from the convergence path towards a Nash equilibrium. This implies that episodes with higher returns tend to be episodes where the agent policies are closer to a joint best response scenario.

The performance of the median performing set of DRL agents is compared to the proposed benchmarks in Table 4.2. As a result of significantly enhancing the utilization of locally available energy, both the average daily import ($\overline{E_{d,+}}$) and export ($\overline{E_{d,-}}$) decline by 21.9% and 84.4%, respectively. Additionally, emergent peak-shaving behavior leads to a lowering of the average daily peak ($\overline{P_{d,+}}$) and valley ($\overline{P_{d,-}}$) by 27.0% and 71.1%, respectively, while the maximum peak (P_+) and minimum valley (P_-) also shrink by 16.0% and 27.0%, respectively. This behavior also results in the smoothing of moment-to-moment community net-load demand, leading to a 26% decrease in the ramping rate ($\overline{R_d}$) and a mitigation of the overall community net-load swing, which reduces the daily load factor ($1 - L_d$) and monthly load factor ($1 - L_m$) by 11.0% and 3.6%, respectively. In summary, ALEX RL significantly mitigates the effects of community-level variability across all measured metrics.

Metric		NoDERMS	IndividualDERMS	ALEX DP	ALEX RL
Average daily import [kWh]	$\overline{E_{d,+}}$	258.54	214.81	202.68	**201.83**
Average daily export [kWh]	$\overline{E_{d,-}}$	-77.48	-26.49	-12.46	**-12.04**
Average daily peak [kW]	$\overline{P_{d,+}}$	25.61	19.95	19.44	**18.69**
Average daily valley [kW]	$\overline{P_{d,-}}$	-16.55	-6.35	**-1.67**	-4.78
Maximum peak [kW]	P_+	49.06	42.37	42.37	**41.22**
Minimum valley [kW]	P_-	-37.86	-36.8	-29.34	**-27.62**
Average daily ramping rate [kW]	$\overline{R_d}$	4.28	2.87	**2.84**	3.15
Average daily load factor	$1 - L_d$	0.73	0.65	**0.64**	0.65
Average monthly load factor	$1 - L_m$	0.82	0.8	**0.78**	0.79

Table 4.2: Summarized metrics for full simulation on CityLearn2022 data set [70] for NoDERMS, IndividualDERMS and ALEX DP and ALEX DRL scenarios. Values for the NoDERMs, IndividualDERMS and ALEX DP are taken out of May et al. [2]. Best values are typeset in bold.

Further comparative analysis demonstrates the cumulative outperformance of the DRL agents against IndividualDERMS and partial outperformance against ALEX DP. Notably, the ramping rate ($\overline{R_d}$) emerges as a sub-performant metric for ALEX RL compared to IndividualDERMS and ALEX DP. Additionally, it is noteworthy that the average daily valley metric ($\overline{P_{d,-}}$) for ALEX RL is significantly higher than ALEX DP, which is somewhat unexpected. While IndividualDERMS and ALEX DP search over a perfect forecast, ALEX RL does not have access to future information and must internally perform some degree of participant net load modeling. As the most short-term volatility-focused metric, the ramping rate ($\overline{R_d}$) is also most sensitive to such misadjustments. The relative disparity in average daily valley ($\overline{P_{d,-}}$) between ALEX RL and ALEX DP may result from a strategic tradeoff, where it is economically safer for the DRL agents to err on the side of selling to the grid than buying from it in the face of an imperfect model. Such a scenario could occur when the market receives significantly more bids than asks in terms of quantity, as the remaining residual load will be settled according to a net-billing scenario.

These results further suggest that ALEX RL compensates for its lack of perfect internal modeling by leveraging its capability to load shift over a longer duration than ALEX DP, resulting in a further decrease in the maximum peak (P_+) and minimum valley (P_-). Therefore, ALEX RL's relative outperformance in terms of bill savings does not necessarily translate to a strict outperformance of ALEX DP in terms of evaluation metrics. Overall, ALEX RL's performance closely aligns with ALEX DP, indicating similar levels of emergent, community-level coordination of DERs. The collective results compellingly demonstrate emergent, community-level variability reduction facilitated by automated participation via DRL agents within a LEM, effectively closing the identified research gap.

In summary, the findings underscore the effectiveness of leveraging DRL agents in LEMs for load optimization. This emphasizes the potential for mitigating variability and optimizing energy consumption at a community level.

4.5 Conclusion

This study explores the automation of participation in economy-driven LEMs through DRL agents.

The rapid proliferation of DERs at the grid edge has led to a significant increase in variability and variance in community net load, posing challenges to electricity grid operability. In response, there has been a growing interest in TE-based DR, facilitated by community LEMs, as a viable solution to align the interests of electricity end-users and grid stakeholders [19, 20, 67]. At the same time, insights from DR system pilots highlight the necessity for automation to ensure robust participation across DR initiatives [8, 9]. In response to the decentralized and distributed nature of this challenge, model-free control approaches, particularly DRL, have emerged as promising candidates [27], fueling the interest in studies investigating the automation of participation in LEMs via DRL methods [13, 14, 17, 18, 49, 84]. While prior research has predominantly focused on socioeconomic metrics and community net load consumption, there remains a gap in demonstrating a clear reduction in variability or variance.

This article addresses the research gap by utilizing a shared experience [89], recurrent PPO [91] algorithm with several modifications [32–34] to train a set of DRL agents within the context of ALEX, an economy-driven LEM where participants aim to selfishly minimize bills without information sharing [1]. The trained DRL agents are compared against benchmark approaches, including a building-level DR strategy and a near-optimal dynamic programming-based solution [2]. Performance is evaluated using a set of metrics capturing net load variance across multiple time horizons, encompassing ramping rate, daily and monthly load factor, peak and average daily import and export. The experiments reveal a clear correlation between relative bill reduction and improvements in the investigated metrics. The trained DRL agents demonstrate promising performance, nearing and, in some instances, surpassing the benchmarks set by the near-optimal approach, while consistently outperforming the building-level DR strategy.

Future research directions should focus on designing more sophisticated DRL algorithms explicitly tailored to the mixed-form stochastic game nature of LEMs like ALEX. The goal is to establish a clearer performance ceiling for such solutions. Additionally, extending this investigation to diverse LEM designs could offer insights into the factors influencing the efficacy of incentivizing desired behaviors within these systems [68, 69].

Chapter 5

Conclusion and Future Research

5.1 Conclusion

In the face of escalating challenges posed by the growing adoption of DERs and the resulting increase in net load variability at the grid-edge, the conventional unidirectional model of grid operations is no longer sufficient. The Smart Grid emerges as a crucial solution to these challenges by focusing on decentralized, intelligent asset integration.

This thesis addresses these issues by proposing the Autonomous Local Energy eXchange (ALEX), demonstrating its operation, and evaluating its efficiency. ALEX, rooted in the principles of TE, is a fully economy-driven LEM automated using DRL agents. The overarching goal is to enhance grid operability and effectively reduce community-level variability by aligning end-user behavior with grid stakeholder objectives while participation and effect are ensured by leveraging decentralized, model-free automation methods.

A reductionist approach is employed to navigate the intricate challenges of such an interconnected system with complex internal dynamics. Key components are isolated through focused experiments, resulting in the formulation of three primary research goals. Each goal is addressed through corresponding journal-grade academic contributions, encapsulated in the chapters of this thesis.

- **Design of a suitable LEM.** Chapter 2, published as "Reinforcement learning-driven local transactive energy market for distributed energy resources" in Energy and AI [1], is motivated by the hypothesis that pricing according to the supply/demand ratio should incentivize the emergent reduction of variability. Such a market is identified from a pool of candidates, delineated by classification criteria for double-auction mechanisms. The insight that ALEX can be formulated as a mixed-form stochastic game, implying the existence of at least one Nash equilibrium, forms the basis of underlying experiments. Beyond its relevance to this specific research goal, this work addresses research gaps acknowledged by other LEM researchers [16, 20] by thoroughly documenting and justifying the process of LEM settlement mechanism design.

- **Establishment of an appropriate benchmarking and evaluation process.** Chapter 3 has been submitted to IEEE Access as "Transactive Local Energy Markets Enable Community-Level Resource Coordination Using Individual Rewards"[2]. It focuses on confirming the central hypothesis regarding emergent, community-level variability reduction within ALEX and develops a benchmarking approach for LEM. The conducted experiments demonstrate the capability of ALEX to enable community-wide coordination of DERs. The benchmarking further shows that ALEX significantly outperforms a set of baseline DR approaches. This is facilitated by a search algorithm that leverages ALEX nature as a mixed-form stochastic game, employing iterative best response and dynamic programming to simulate a set of near-optimal policies close to a Nash equilibrium. In the process, this chapter also addresses broader research gaps in current LEM literature [68, 69] such as the general lack of comparison platforms. This is achieved through using an open-source dataset and benchmarking via metrics that evaluate LEM performance regarding variability instead of socioeconomic considerations.
- **Development of an adequate DRL algorithm and training routine.**

The conclusive Chapter 4, submitted to Energy and AI as “Decentralized Coordination of Distributed Energy Resources through Local Energy Markets and Deep Reinforcement Learning” [3], marks the conclusion of the thesis by training DRL agents to near-optimal performance on ALEX. Related work combining DRL agents with LEM [14, 17, 18, 49, 84] largely evaluates their agents shallowly on socioeconomic or self-sufficiency-related metrics, lacking a convincing showcase of community-level variability reduction. By demonstrating the emergence of variability reducing coordination across a community in this setting, this study closes a significant research gap. The utilized DRL algorithm is based on PPO [31] and augmented with several general algorithmic improvements [33, 89, 94]. ALEX is formulated as an MDP with continuous action and observation space, emphasizing observation and policy design to accelerate learning. The agents are trained on the same dataset employed in Chapter 3, demonstrating performance very close to the established near-optimal policies while consistently outperforming all other baselines. This demonstrates the ability of the proposed algorithm to produce a convergent set of policies close to a Nash equilibrium, even without information sharing between agents and access to future information.

In summary, this thesis significantly advances the state of the art of indirect DR by proposing and demonstrating ALEX. The decentralized, autonomous nature of ALEX positions itself as suitable for addressing the challenges posed by the growing adoption of DERs in the shape of a system that aligns itself with the Smart Grid’s tenets of intelligent integration of all system participants towards ensuring efficient and reliable grid operation.

5.2 Future Work

The contributions presented in this thesis, including the development and demonstration of ALEX, have laid a foundation for further exploration and sparked commercialization efforts. Implementation-focused future work considerations revolving around hardware implementation, communication modeling for deployment, and cybersecurity are better suited to this commercialization environment. Additionally, the improvements made to PPO within Chapter 4 are general but not exhaustive. Further advancements, such as prioritized replay buffers [95], regularization, and intrinsic motivation for exploration [96], should be explored for direct algorithm improvement within the commercialization context. With that in mind, the presented future work suggestions maintain a software focus and consider integration into the academic environment.

The convergence of participants to optimal prices has been demonstrated in bandit experiments in Chapter 2, and further confirmed in preliminary experiments for Chapter 4. However, the near-optimal participants within ALEX with the originally envisioned multi-modal action space encompassing bid and ask quantities, prices, and DER control have not been demonstrated yet. Exploring DRL agents with this multi-modal action space represents an ambitious research avenue. Such investigations might be better suited towards evaluation within the computer science context, focusing on more contained multi-agent systems in mixed-form stochastic games with multi-modal action spaces. This research would have a clear pathway toward publication, whilst simultaneously holding significant potential for follow-up integration into the commercialization pathway.

The DRL algorithm proposed and utilized in Chapter 4 can be enhanced on several fronts. Currently, it does not explicitly account for the mixed-form stochastic game nature of ALEX. A more sophisticated approach considering this characteristic could potentially outperform the existing algorithm. The field of multi-agent DRL is actively researched, and exploring more advanced techniques in a larger project that benchmarks the proposed algorithm on other applications holds academic promise.

Another intriguing direction for future research involves a comprehensive exploration of different double auction markets and other settlement mechanism intricacies. Preliminary studies may be based on a bandit RL setup akin to the described experiments in Chapter 2. Still, the research should aim to leverage a search process similar to the one implemented in Chapter 3 to grade market mechanisms based on their direct effects on community net-load variability. The wide variety of LEMs proposed in recent literature [19–21, 67] adds further motivation for such investigations. Given the landscape of LEM literature, such research has a promising pathway to publication within the academic context and also possible commercialization implications.

Extending the benchmarking efforts initiated in Chapters 3 and 4 to include other DR methodologies would be valuable for the broader DR field. Reproduction studies benchmarking various DR approaches using common datasets and metrics could elevate future research. However, such an endeavor might be challenging to publish as a standalone contribution and would be better suited for an advanced course format.

Bibliography

[1] S. Zhang, D. May, M. Gül, and P. Musilek, “Reinforcement learning-driven local transactive energy market for distributed energy resources,” *Energy and AI*, vol. 8, p. 100 150, 2022, ISSN: 2666-5468. DOI: https://doi.org/10.1016/j.egyai.2022.100150. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666546822000118.

[2] D. May and P. Musilek, “Transactive local energy markets enable community-level resource coordination using individual rewards,” *arXiv preprint arXiv:2403.15617*, 2024. DOI: https://doi.org/10.48550/arXiv.2403.15617. [Online]. Available: https://arxiv.org/abs/2403.15617.

[3] D. May, M. Taylor, and P. Musilek, “Decentralized coordination of distributed energy resources through local energy markets and deep reinforcement learning,” *Energy and AI*, 2024. DOI: TBD.

[4] I. R. E. Agency, “Global energy transformation: A roadmap to 2050 (2019 edition),” International Renewable Energy Agency, Tech. Rep., 2019. [Online]. Available: https://www.irena.org/publications/2019/Apr/Global-energy-transformation-A-roadmap-to-2050-2019Edition.

[5] I. E. Agency, *World Energy Outlook 2022*. 2022, p. 524. DOI: https://doi.org/https://doi.org/10.1787/3a469970-en. [Online]. Available: https://www.oecd-ilibrary.org/content/publication/3a469970-en.

[6] M. Fera, R. Macchiaroli, R. Iannone, S. Miranda, and S. Riemma, “Economic evaluation model for the energy demand response,” *Energy*, vol. 112, pp. 457–468, 2016, ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2016.06.123. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544216308970.

[7] J. Aghaei and M.-I. Alizadeh, “Demand response in smart electricity grids equipped with renewable energy sources: A review,” *Renewable and Sustainable Energy Reviews*, vol. 18, pp. 64–72, 2013, ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2012.09.019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032112005205.

[8] F. E. R. Commission, “Assessment of demand response and advanced metering,” Federal Energy Regulatory Commission, Tech. Rep., 2015. [Online]. Available: https://www.ourenergypolicy.org/wp-content/uploads/2015/12/demand-response.pdf.

[9] S. Chen and C.-C. Liu, “From demand response to transactive energy: State of the art,” *Journal of Modern Power Systems and Clean Energy*, vol. 5, no. 1, pp. 10–19, 2017. DOI: 10.1007/s40565-016-0256-x.

[10] R. B. Melton, “Gridwise transactive energy framework,” The GridWise Architecture Council, Tech. Rep., version 1.1. 2019. [Online]. Available: https://gridwiseac.org/pdfs/pnnl_22946_gwac_te_framework_july_2019_v1_1.pdf.

[11] H.-M. Chung, S. Maharjan, Y. Zhang, and F. Eliassen, “Distributed deep reinforcement learning for intelligent load scheduling in residential smart grids,” *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2752–2763, 2021. DOI: https://doi.org/10.1109/TII.2020.3007167.

[12] Q. Zhang, K. Dehghanpour, Z. Wang, F. Qiu, and D. Zhao, “Multi-agent safe policy learning for power management of networked microgrids,” *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 1048–1062, 2021. DOI: https://doi.org/10.1109/TSG.2020.3034827.

[13] Y. Ye, Y. Tang, H. Wang, X.-P. Zhang, and G. Strbac, “A scalable privacy-preserving multi-agent deep reinforcement learning approach for large-scale peer-to-peer transactive energy trading,” *IEEE Transactions on Smart Grid*, vol. 12, no. 6, pp. 5185–5200, 2021. DOI: https://doi.org/10.1109/TSG.2021.3103917.

[14] Y. Ye, D. Papadaskalopoulos, Q. Yuan, Y. Tang, and G. Strbac, “Multi-agent deep reinforcement learning for coordinated energy trading and flexibility services provision in local electricity markets,” *IEEE Transactions on Smart Grid*, vol. 14, no. 2, pp. 1541–1554, 2023. DOI: https://doi.org/10.1109/TSG.2022.3149266.

[15] S. Bose, E. Kremers, E. M. Mengelkamp, J. Eberbach, and C. Weinhardt, “Reinforcement learning in local energy markets,” *Energy Informatics*, vol. 4, no. 1, p. 7, May 2021, ISSN: 2520-8942. DOI: https://doi.org/10.1186/s42162-021-00141-z.

[16] E. Mengelkamp, J. Gärttner, and C. Weinhardt, “Intelligent agent strategies for residential customers in local electricity markets,” in *Proceedings of the Ninth International Conference on Future Energy Systems*, ser. e-Energy ’18, Karlsruhe, Germany: Association for Computing Machinery, 2018, 97–107, ISBN: 9781450357678. DOI: https://doi.org/10.1145/3208903.3208907. [Online]. Available: https://doi.org/10.1145/3208903.3208907.

[17] S. Zhou, Z. Hu, W. Gu, M. Jiang, and X.-P. Zhang, “Artificial intelligence based smart energy community management: A reinforcement learning approach,” *CSEE Journal of Power and Energy Systems*, vol. 5, no. 1, pp. 1–10, 2019. DOI: https://doi.org/10.17775/CSEEJPES.2018.00840.

[18] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, and C. S. Lai, “A multi-agent reinforcement learning-based data-driven method for home energy management,” *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3201–3211, 2020. DOI: https://doi.org/10.1109/TSG.2020.2971427.

[19] T. Capper *et al.*, "Peer-to-peer, community self-consumption, and transactive energy: A systematic literature review of local energy market models," *Renewable and Sustainable Energy Reviews*, vol. 162, p. 112 403, 2022, ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2022.112403. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032122003112.

[20] E. Mengelkamp, J. Diesing, and C. Weinhardt, "Tracing local energy markets: A literature review:" *it - Information Technology*, vol. 61, no. 2-3, pp. 101–110, 2019. DOI: https://doi.org/10.1515/itit-2019-0016.

[21] W. Tushar *et al.*, "Peer-to-peer energy systems for connected communities: A review of recent advances and emerging challenges," *Applied Energy*, vol. 282, p. 116 131, 2021, ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2020.116131. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261920315464.

[22] Y. Lin, J. McPhee, and N. L. Azad, "Comparison of deep reinforcement learning and model predictive control for adaptive cruise control," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 2, pp. 221–231, 2021. DOI: 10.1109/TIV.2020.3012947.

[23] T. A. Badgwell, J. H. Lee, and K.-H. Liu, "Reinforcement learning – overview of recent progress and implications for process control," in *13th International Symposium on Process Systems Engineering (PSE 2018)*, ser. Computer Aided Chemical Engineering, M. R. Eden, M. G. Ierapetritou, and G. P. Towler, Eds., vol. 44, Elsevier, 2018, pp. 71–85. DOI: https://doi.org/10.1016/B978-0-444-64241-7.50008-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780444642417500082.

[24] B. Singh, R. Kumar, and V. P. Singh, "Reinforcement learning in robotic applications: A comprehensive survey," *Artificial Intelligence Review*, pp. 1–46, 2022. DOI: https://doi.org/10.1007/s10462-021-09997-9.

[25] J. Shin, T. A. Badgwell, K.-H. Liu, and J. H. Lee, "Reinforcement learning – overview of recent progress and implications for process control," *Computers and Chemical Engineering*, vol. 127, pp. 282–294, 2019, ISSN: 0098-1354. DOI: https://doi.org/10.1016/j.compchemeng.2019.05.029. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0098135419300754.

[26] R. Sutton, "The bitter lesson," *Incomplete Ideas (blog)*, vol. 13, no. 1, 2019.

[27] J. R. Vázquez-Canteli and Z. Nagy, "Reinforcement learning for demand response: A review of algorithms and modeling techniques," *Applied Energy*, vol. 235, pp. 1072–1089, 2019, ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2018.11.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261918317082.

[28] D. Cao *et al.*, "Reinforcement learning and its applications in modern power and energy systems: A review," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1029–1042, 2020. DOI: 10.35833/MPCE.2020.000552.

[29] A. Perera and P. Kamalaruban, "Applications of reinforcement learning in energy systems," *Renewable and Sustainable Energy Reviews*, vol. 137, p. 110 618, 2021, ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2020.110618. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032120309023.

[30] B. Baker *et al.*, "Emergent tool use from multi-agent autocurricula," *CoRR*, vol. abs/1909.07528, 2019. arXiv: 1909.07528. [Online]. Available: http://arxiv.org/abs/1909.07528.

[31] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017. DOI: https://doi.org/10.48550/arXiv.1707.06347. arXiv: 1707.06347. [Online]. Available: http://arxiv.org/abs/1707.06347.

[32] S. Kapturowski, G. Ostrovski, J. Quan, R. Munos, and W. Dabney, "Recurrent experience replay in distributed reinforcement learning," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:59345798.

[33] A. Ilyas *et al.*, "Are deep policy gradient algorithms truly policy gradient algorithms?" *CoRR*, vol. abs/1811.02553, 2018. arXiv: 1811.02553. [Online]. Available: http://arxiv.org/abs/1811.02553.

[34] M. Andrychowicz *et al.*, *What matters in on-policy reinforcement learning? a large-scale empirical study*, 2020. DOI: https://doi.org/10.48550/arXiv.2006.05990. arXiv: 2006.05990 [cs.LG].

[35] A. Fattahi and M. Deihimi, "A review of demand-side management: Reconsidering theoretical framework," *Renewable and Sustainable Energy Reviews*, vol. 80, pp. 367–379, Dec. 2017. DOI: https://doi.org/10.1016/j.rser.2017.05.207.

[36] B.-G. Kim, Y. Zhang, M. van der Schaar, and J.-W. Lee, "Dynamic pricing and energy consumption scheduling with reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2187–2198, 2016. DOI: https://doi.org/10.1109/TSG.2015.2495145.

[37] R. Lu, S. Hong, and X. Zhang, "A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach," *Applied Energy*, vol. 220, pp. 220–230, Jun. 2018. DOI: https://doi.org/10.1016/j.apenergy.2018.03.072.

[38] A. Meyabadi and M. Deihimi, "A review of demand-side management: Reconsidering theoretical framework," *Renewable and Sustainable Energy Reviews*, vol. 80, pp. 367–379, 2017, ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2017.05.207.

[39] O. Abrishambaf, F. Lezama, P. Faria, and Z. Vale, "Towards transactive energy systems: An analysis on current trends," *Energy Strategy Reviews*, vol. 26, p. 100 418, 2019, ISSN: 2211-467X. DOI: https://doi.org/10.1016/j.esr.2019.100418. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2211467X19301105.

[40] M. Yu, R. Lu, and S. Hong, "A real-time decision model for industrial load management in a smart grid," *Applied Energy*, vol. 183, Dec. 2016. DOI: https://doi.org/10.1016/j.apenergy.2016.09.021.

[41] X. Huang, S. H. Hong, and Y. Li, "Hour-ahead price based energy management scheme for industrial facilities," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 6, pp. 2886–2898, 2017. DOI: https://doi.org/10.1109/TII.2017.2711648.

[42] R. de Sá Ferreira, L. A. Barroso, P. R. Lino, M. M. Carvalho, and P. Valenzuela, "Time-of-use tariff design under uncertainty in price-elasticities of electricity demand: A stochastic optimization approach," *IEEE Transactions on Smart Grid*, vol. 4, no. 4, pp. 2285–2295, 2013. DOI: https://doi.org/10.1109/TSG.2013.2241087.

[43] Y. Liu, D. Zhang, C. Deng, and X. Wang, "Deep reinforcement learning approach for autonomous agents in consumer-centric electricity market," in *2020 5th IEEE International Conference on Big Data Analytics (ICBDA)*, 2020, pp. 37–41. DOI: https://doi.org/10.1109/ICBDA49040.2020.9099946.

[44] D. Forfia, M. Knight, and R. Melton, "The view from the top of the mountain: Building a community of practice with the gridwise transactive energy framework," *IEEE Power and Energy Magazine*, vol. 14, no. 3, pp. 25–33, 2016. DOI: 10.1109/MPE.2016.2524961.

[45] M. Pilz and L. Al-Fagih, "Recent advances in local energy trading in the smart grid based on game-theoretic approaches," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 1363–1371, 2019. DOI: https://doi.org/10.1109/TSG.2017.2764275.

[46] M. Khorasany, Y. Mishra, and G. Ledwich, "Market framework for local energy trading: A review of potential designs and market clearing approaches," *IET Generation, Transmission & Distribution*, vol. 12, no. 22, pp. 5899–5908, 2018. DOI: https://doi.org/10.1049/iet-gtd.2018.5309.

[47] E. Mengelkamp, P. Staudt, J. Garttner, and C. Weinhardt, "Trading on local energy markets: A comparison of market designs and bidding strategies," in *2017 14th International Conference on the European Energy Market (EEM)*, 2017, pp. 1–6. DOI: 10.1109/EEM.2017.7981938.

[48] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Second. The MIT Press, 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html.

[49] H. Zang and J. Kim, "Reinforcement learning based peer-to-peer energy trade management using community energy storage in local energy market," *Energies*, vol. 14, no. 14, 2021, ISSN: 1996-1073. DOI: https://doi.org/10.3390/en14144131. [Online]. Available: https://www.mdpi.com/1996-1073/14/14/4131.

[50] E. Foruzan, L.-K. Soh, and S. Asgarpoor, “Reinforcement learning approach for optimal distributed energy management in a microgrid,” *IEEE Transactions on Power Systems*, vol. 33, no. 5, pp. 5749–5758, 2018. DOI: https://doi.org/10.1109/TPWRS.2018.2823641.

[51] S. Zhou, Z. Hu, W. Gu, M. Jiang, and X.-P. Zhang, “Artificial intelligence based smart energy community management: A reinforcement learning approach,” *CSEE Journal of Power and Energy Systems*, vol. 5, no. 1, pp. 1–10, 2019. DOI: https://doi.org/10.17775/CSEEJPES.2018.00840.

[52] T. Chen and W. Su, “Local energy trading behavior modeling with deep reinforcement learning,” *IEEE Access*, vol. 6, pp. 62 806–62 814, 2018. DOI: https://doi.org/10.1109/ACCESS.2018.2876652.

[53] S. Sunder and D. Gode, “Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality,” *Journal of Political Economy*, vol. 101, pp. 119–37, Feb. 1993. DOI: https://doi.org/10.1086/261868.

[54] T. Chen and W. Su, “Indirect customer-to-customer energy trading with reinforcement learning,” *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 4338–4348, 2019. DOI: https://doi.org/10.1109/TSG.2018.2857449.

[55] O. Jogunola, Y. Tsado, B. Adebisi, and R. Nawaz, “Trading strategy in a local energy market, a deep reinforcement learning approach,” in *2021 IEEE Electrical Power and Energy Conference (EPEC)*, 2021, pp. 347–352. DOI: https://doi.org/10.1109/EPEC52095.2021.9621459.

[56] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, *Prioritized experience replay*, 2016. DOI: https://doi.org/10.48550/arXiv.1511.05952. arXiv: 1511.05952 [cs.LG].

[57] I. Erev and A. E. Roth, “Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria,” *The American Economic Review*, vol. 88, no. 4, pp. 848–881, 1998, ISSN: 00028282. [Online]. Available: http://www.jstor.org/stable/117009.

[58] J. Nicolaisen, V. Petrov, and L. Tesfatsion, “Market power and efficiency in a computational electricity market with discriminatory double-auction pricing,” *IEEE Transactions on Evolutionary Computation*, vol. 5, no. 5, pp. 504–523, 2001. DOI: 10.1109/4235.956714.

[59] D. J. Harrold, J. Cao, and Z. Fan, “Data-driven battery operation for energy arbitrage using rainbow deep reinforcement learning,” *Energy*, vol. 238, p. 121 958, 2022, ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2021.121958. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544221022064.

[60] S. Lee and D.-H. Choi, “Dynamic pricing and energy management for profit maximization in multiple smart electric vehicle charging stations: A privacy-preserving deep reinforcement learning approach,” *Applied Energy*, vol. 304, p. 117 754, Dec. 2021. DOI: https://doi.org/10.1016/j.apenergy.2021.117754.

[61] F. Oliehoek and C. Amato, *A Concise Introduction to Decentralized POMDPs*. Jan. 2016, ISBN: 978-3-319-28927-4. DOI: https://doi.org/10.1007/978-3-319-28929-8.

[62] R. B. Myerson and M. A. Satterthwaite, "Efficient mechanisms for bilateral trading," *Journal of Economic Theory*, vol. 29, no. 2, pp. 265–281, 1983, ISSN: 0022-0531. DOI: https://doi.org/10.1016/0022-0531(83)90048-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022053183900480.

[63] J. Hu and M. P. Wellman, "Nash q-learning for general-sum stochastic games," *J. Mach. Learn. Res.*, vol. 4, no. null, 1039–1069, 2003, ISSN: 1532-4435.

[64] S. Zhang, *Trex-publication-resources*, https://github.com/sd-zhang/publications-resources, 2021.

[65] D. Chen and D. Irwin, "Sundance: Black-box behind-the-meter solar disaggregation," in *Proceedings of the Eighth International Conference on Future Energy Systems*, ser. e-Energy '17, Association for Computing Machinery, 2017, 45–55, ISBN: 9781450350365. DOI: https://doi.org/10.1145/3077839.3077848. [Online]. Available: https://doi.org/10.1145/3077839.3077848.

[66] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, and J. Albrecht, "Smart*: An open data set and tools for enabling research in sustainable homes," *Proc. SustKDD.*, Jan. 2012.

[67] V. Dudjak *et al.*, "Impact of local energy markets integration in power systems layer: A comprehensive review," *Applied Energy*, vol. 301, p. 117 434, 2021, ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2021.117434. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261921008266.

[68] D. Kiedanski, D. Kofman, P. Maillé, and J. Horta, "Misalignments of objectives in demand response programs: A look at local energy markets," in *2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 2020, pp. 1–7. DOI: https://doi.org/10.1109/SmartGridComm47815.2020.9302939.

[69] D. Papadaskalopoulos and G. Strbac, "Nonlinear and randomized pricing for distributed management of flexible loads," *IEEE Transactions on Smart Grid*, vol. 7, no. 2, pp. 1137–1146, 2016. DOI: 10.1109/TSG.2015.2437795.

[70] K. Nweye, S. Siva, and G. Z. Nagy, *The CityLearn Challenge 2022*, version V1, 2023. DOI: https://doi.org/10.18738/T8/0YLJ6Q.

[71] K. Nweye, S. Sankaranarayanan, and Z. Nagy, "Merlin: Multi-agent offline and transfer learning for occupant-centric operation of grid-interactive communities," *Applied Energy*, vol. 346, p. 121 323, 2023, ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2023.121323. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261923006876.

[72] E. Mengelkamp, S. Bose, E. Kremers, J. Eberbach, B. Hoffmann, and C. Weinhardt, "Increasing the efficiency of local energy markets through residential demand response," *Energy Informatics*, vol. 1, pp. 1–18, 2018. DOI: https://doi.org/10.1186/s42162-018-0017-3. [Online]. Available: https://doi.org/10.1186/s42162-018-0017-3.

[73] J. R. Vazquez-Canteli, G. Henze, and Z. Nagy, "Marlisa: Multi-agent reinforcement learning with iterative sequential action selection for load shaping of grid-interactive connected buildings," in *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, ser. BuildSys '20, Association for Computing Machinery, 2020, 170–179, ISBN: 9781450380614. DOI: https://doi.org/10.1145/3408308.3427604.

[74] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, and C. S. Lai, "A multi-agent reinforcement learning-based data-driven method for home energy management," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3201–3211, 2020. DOI: https://doi.org/10.1109/TSG.2020.2971427.

[75] S. Zhang and P. Musilek, "The impact of battery storage on power flow and economy in an automated transactive energy market," *Energies*, vol. 16, no. 5, 2023, ISSN: 1996-1073. DOI: https://doi.org/10.3390/en16052251. [Online]. Available: https://www.mdpi.com/1996-1073/16/5/2251.

[76] R. Bellman, "A markovian decision process," *Journal of Mathematics and Mechanics*, vol. 6, no. 5, pp. 679–684, 1957, ISSN: 00959057, 19435274. [Online]. Available: http://www.jstor.org/stable/24900506 (visited on 07/21/2023).

[77] R. Bellman, *Dynamic Programming*, 1st ed. Princeton, NJ, USA: Princeton University Press, 1957.

[78] R. Coulom, "Efficient selectivity and backup operators in monte-carlo tree search," in *Computers and Games*, H. J. van den Herik, P. Ciancarini, and H. H. L. M. J. Donkers, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 72–83, ISBN: 978-3-540-75538-8.

[79] K. Kok and S. Widergren, "A society of devices: Integrating intelligent distributed resources with transactive energy," *IEEE Power and Energy Magazine*, vol. 14, no. 3, pp. 34–45, 2016. DOI: https://doi.org/10.1109/MPE.2016.2524962.

[80] A. O'Connell, J. Taylor, J. Smith, and L. Rogers, "Distributed energy resources takes center stage: A renewed spotlight on the distribution planning process," *IEEE Power and Energy Magazine*, vol. 16, no. 6, pp. 42–51, 2018. DOI: https://doi.org/10.1109/MPE.2018.2862439.

[81] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015. DOI: https://doi.org/10.1038/nature14236.

[82] D. Silver *et al.*, “A general reinforcement learning algorithm that masters chess, shogi, and go through self-play,” *Science*, vol. 362, no. 6419, pp. 1140–1144, 2018. DOI: https://doi.org/10.1126/science.aar6404. eprint: https://www.science.org/doi/pdf/10.1126/science.aar6404. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aar6404.

[83] OpenAI *et al.*, “Dota2 with large scale deep reinforcement learning,” 2019. DOI: https://doi.org/10.48550/arXiv.1912.06680. arXiv: 1912.06680 `[cs.LG]`.

[84] T. Chen and S. Bu, “Realistic peer-to-peer energy trading model for microgrids using deep reinforcement learning,” in *2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*, 2019, pp. 1–5. DOI: https://doi.org/10.1109/ISGTEurope.2019.8905731.

[85] N. Liu, X. Yu, C. Wang, C. Li, L. Ma, and J. Lei, “Energy-sharing model with price-based demand response for microgrids of peer-to-peer prosumers,” *IEEE Transactions on Power Systems*, vol. 32, no. 5, pp. 3569–3583, 2017. DOI: https://doi.org/10.1109/TPWRS.2017.2649558.

[86] F. Lezama, J. Soares, P. Hernandez-Leal, M. Kaisers, T. Pinto, and Z. Vale, “Local energy markets: Paving the path toward fully transactive energy systems,” *IEEE Transactions on Power Systems*, vol. 34, no. 5, pp. 4081–4088, 2019. DOI: 10.1109/TPWRS.2018.2833959.

[87] R. Ghorani, M. Fotuhi-Firuzabad, and M. Moeini-Aghtaie, “Optimal bidding strategy of transactive agents in local energy markets,” *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5152–5162, 2019. DOI: https://doi.org/10.1109/TSG.2018.2878024.

[88] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, “High-dimensional continuous control using generalized advantage estimation,” *arXiv preprint arXiv:1506.02438*, 2015. DOI: https://doi.org/10.48550/arXiv.1506.02438.

[89] F. Christianos, L. Schäfer, and S. Albrecht, “Shared experience actor-critic for multi-agent reinforcement learning,” vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., pp. 10 707–10 717, 2020. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/7967cc8e3ab559e68cc944c44b1cf3e8-Paper.pdf.

[90] J. Schulman, “The nuts and bolts of deep rl research,” in *NIPS Deep RL Workshop*, 2016. [Online]. Available: http://joschu.net/docs/nuts-and-bolts.pdf.

[91] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, “Stable-baselines3: Reliable reinforcement learning implementations,” *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-1364.html.

[92] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, 1735–1780, 1997, ISSN: 0899-7667. DOI: https://doi.org/10.1162/neco.1997.9.8.1735. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735.

[93] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 1861–1870. [Online]. Available: https://proceedings.mlr.press/v80/haarnoja18b.html.

[94] K. Gupta, “What matters in recurrent ppo for long episodic and continuing partially observable tasks,” [Online]. Available: https://kshitijkg.github.io/data/RecurrentPPO_Report.pdf.

[95] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, “Prioritized experience replay,” *arXiv preprint arXiv:1511.05952*, 2015. DOI: https://doi.org/10.48550/arXiv.1511.05952.

[96] T. D. Kulkarni, K. Narasimhan, A. Saeedi, and J. Tenenbaum, “Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation,” vol. 29, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/f442d33fa06832082290ad8544a8da27-Paper.pdf.

[97] D. Silver *et al.*, “Mastering the game of go with deep neural networks and tree search,” *Nature*, vol. 529, pp. 484–489, Jan. 2016. DOI: https://doi.org/10.1038/nature16961.

[98] O. Vinyals *et al.*, “Grandmaster level in StarCraft II using multi-agent reinforcement learning,” *Nature*, vol. 575, no. 7782, 2019, Number: 7782 Publisher: Nature Publishing Group. DOI: https://doi.org/10.1038/s41586-019-1724-z. (visited on 02/24/2021).

[99] D. Arrachequesne, *Socket.io*, 2021. [Online]. Available: https://github.com/socketio/socket.io (visited on 02/24/2021).

[100] EPRI, *Epri distribution system simulator*, 2021. [Online]. Available: https://sourceforge.net/projects/electricdss/ (visited on 02/24/2021).

[101] D. Krishnamurthy, *Opendssdirect.py*, https://github.com/dss-extensions/OpenDSSDirect.py, 2017.

[102] L. Tesfatsion, “Agent-based computational economics: Growing economies from the bottom up.,” *Artificial Life*, vol. 8, no. 1, pp. 55–82, 2002. DOI: 10.1162/106454602753694765.

[103] D. Friedman, *The double auction market: institutions, theories, and evidence.* Routledge, 2018.

[104] D. Friedman, “A simple testable model of double auction markets,” *Journal of Economic Behavior and Organization*, vol. 15, no. 1, pp. 47–70, 1991, ISSN: 0167-2681. DOI: https://doi.org/10.1016/0167-2681(91)90004-H. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016726819190004H.

[105] M. Pleines, M. Pallasch, F. Zimmer, and M. Preuss, *Generalization, mayhems and limits in recurrent proximal policy optimization*, 2022. DOI: https://doi.org/10.48550/arXiv.2205.11104. arXiv: 2205.11104 `[cs.LG]`.

[106] S. Huang, R. F. J. Dossa, A. Raffin, A. Kanervisto, and W. Wang, "The 37 implementation details of proximal policy optimization," *The ICLR Blog Track 2023*, 2022. [Online]. Available: https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/.

Appendix A: Appendices for Chapter 2

A.1 Reinforcement Learning

There have been numerous approaches used to address the optimization problems inherent to demand side management [9, 35, 38, 39]including mixed-integer programming, stochastic programming, and dynamic programming. After RL demonstrated great competence in partially observable, stochastic game environments [83, 97, 98], this model-free, sequence-oriented, semi-supervised machine learning framework has also gained popularity as a control method for DR [27]. The learned policy can substitute the solution of the equivalent optimization problem at each time step. As a result, RL approaches can be more computationally efficient at scale, when compared to conventional optimization methods.

In the RL setting, illustrated in Figure A.1, an agent learns to maximize the return G by interacting with its environment through actions a while receiving observations of the environmental state s. G is commonly defined as the expected, discounted cumulant of future reward R

$$G_t = \sum_{i=t}^{T} \gamma^{(i-t)} R_i, \ \forall \gamma \in [0...1], \tag{A.1}$$

where γ is the discount factor. The expected value of G_t, given the current state s_t

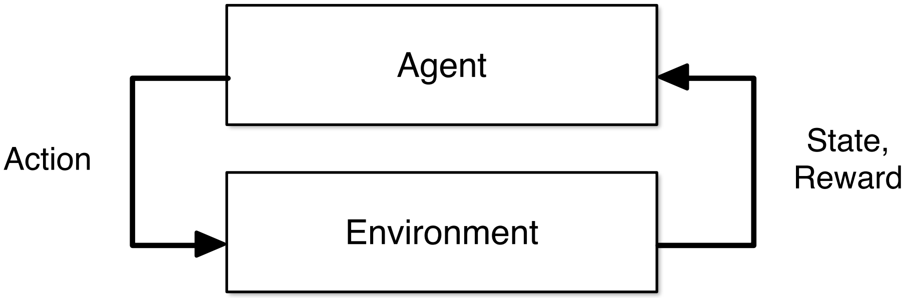

Figure A.1: Reinforcement Learning Setting

or the current state-action tuple (s_t, a_t), is referred to as the state value

$$V(s_t) = \mathbb{E}\big(G_t|_{s_t,\pi(s_t)}\big), \tag{A.2}$$

or action value

$$Q(s_t, a_t) = \mathbb{E}\big(G_t|_{s_t,a_t}\big), \tag{A.3}$$

respectively. An RL agent acts according to a policy π

$$\pi : S \times A \rightarrow [0...1]. \tag{A.4}$$

Policy is a (probabilistic) mapping of the state space S on action space A, i.e.

$$\sum_a \pi(a, s_t) = 1. \tag{A.5}$$

This allows the definition of the state value V as action value Q weighted by π

$$V(s_t) = \frac{1}{n_a} \sum_a \pi(a, s_t) Q(s_t, a). \tag{A.6}$$

This system of equations (A.1-A.6) is sufficient to broadly classify all RL algorithms along two axes: the learned function and the relation between the target and behavior policy. According to the learned function, RL algorithms can be classified as policy-gradient methods and value-based methods. The policy-gradient methods directly learn the policy π, while the value-based methods learn estimations for either V or Q, and employ a fixed mapping of these values to π. RL algorithms can also be classified into on-policy and off-policy methods, by comparing their target and exploratory behavior. An on-policy RL algorithm explores the environment with the same policy that is optimized, while an off-policy algorithm explores the environment with a behavioral policy $b \neq \pi$.

Internally, RL algorithms often employ function approximation techniques to perform the mapping of the state space S to the learned target, and therefore the return G. Historically, tabular encoding was commonly used while currently deep artificial neural networks are currently most popular. As a framework, RL is independent of the choice of state estimator. Currently, a very popular choice is the use of deep artificial neural networks. Historically, other function approximation techniques have also been used, such as tabular encoding.

The algorithm used in this paper is Q-learning, a well-established, value-based, off-policy algorithm. It learns the greedy policy, a deterministic policy that always picks a corresponding to the largest Q, by following behavioral policy b. A popular choice for b is the ϵ-greedy policy, which takes a random action with probability ϵ and otherwise follows the greedy policy. The corresponding learning rule can be written as

$$\begin{aligned} Q^{\text{updated}}(a_t, s_t) \leftarrow (1-\alpha)Q(a_t, s_t) + \\ \alpha\Big(R_t + \gamma \max_a\big(Q(a, s_{t+1})\big)\Big), \end{aligned} \tag{A.7}$$

where α is the learning rate.

Q-learning is a relatively well-understood RL algorithm, with strong convergence criteria for the tabular function approximation case, as long as both ϵ and α are annealed towards 0 at infinity. It is also the most common algorithm in the related literature reviewed in section 2.2.1.

A.2 Net Billing

This appendix clarifies the distinction between net billing and net metering. In certain jurisdictions, such as Alberta, Canada, the electricity market is ”unbundled”. In simple terms, electric utilities are only in charge of building and operating the infrastructure (wires), and a multitude of retailers (which cannot be the same entity as the electric utility) are allowed to sell electricity to end users, with almost complete freedom to set the rate of electricity. Customer bills are therefore also separated into two main components: infrastructure (transmission and distribution, or T&D fees, which can have a fixed component and a variable component), and energy. Under net metering, any electricity that flows into the meter (loads) incurs both energy and variable T&D costs, and any electricity that flows out of the meter (generation) has both energy and T&D costs deducted, either as credits or cashback. Net billing is the same for loads, but only the energy component is deducted for generation. One way to avoid this infrastructure cost is to install both the solar panel and a battery behind the meter to minimize the amount of energy flowing out of the meter. The advantage of net billing is the socialized cost of infrastructure, which is more evenly divided amongst all customers. In contrast, net metering tends to shift these costs onto the segment of the population who cannot afford their own solar (this is a commonly known and often criticized problem). The disadvantage of net billing is that the return on investment (ROI) can be significantly longer due to less bill deductions. From this perspective, net billing is a more fair baseline. It also provides more opportunities for community-based energy management, such as through local energy markets like ALEX.

A.3 T-REX

A.3.1 System Architecture

The major limitations in deploying any TE technology at a scale are communication infrastructure and computational power. This is especially true for the distribution system, where the amount of data that needs to be collected and processed for TE is orders of magnitude greater compared to the transmission system. Furthermore, the necessary infrastructure, such as SCADA, private fiber networks, voltage sensors, current sensors, etc., is typically unavailable to the distribution system and would be prohibitively expensive to retrofit.

The T-REX architecture is therefore designed around the least expensive way to implement and scale TE technology. This means that inexpensive, low bandwidth, long-range wireless mesh networks, such as LoRaWAN, can be used for reliable communications. Computing devices should also be distributed so that the total computational power of the network scales with the number of TE clients. Figure A.2 shows the simplified architecture diagram of this approach.

A.3.2 Data Fabric

T-REX is built upon `socket.io` [99] as the system foundation. This guarantees compatibility, scalability, reliability, and deployability. When using T-REX in simulation mode, the asynchronous, highly parallel design provides true system-wide randomness and eliminates the need for pseudo-random sequence queues that are typically required for contemporary TE simulators. Standard networking performance and penetration testing techniques can be readily used to evaluate the performance and cybersecurity aspects of the TE systems designed with T-REX.

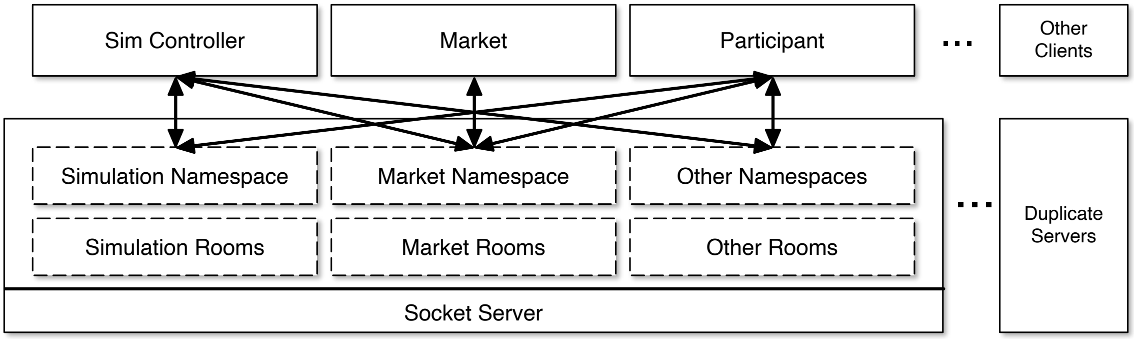

Figure A.2: Simplified T-REX V3 Architecture Diagram

A.3.3 Clients

The functional modules of T-REX are built as `socket.io` clients. As in the deployment case, interaction between modules is facilitated by passing messages using the `socket.io` API. Although designers are free to use payloads of any permissible size, format, and endpoints, care should be taken to preserve genericity and minimize bandwidth usage. There have been three main classes of clients implemented, as shown in the architecture diagram and described below:

- Participant modules, which are in charge of energy trading and managing energy resources that are directly accessible. Participants are, for example, households and self-driving EVs.
- Non-participant modules, e.g., the TE market. The market facilitates the discovery and exchange of energy between participants.
- Simulation-only modules, e.g., the simulation controller, or a powerflow calculation module. The simulation controller augments the deployment environment to form a simulation model. It can also perform advanced functions such as training curricula for ML applications.

With a few restrictions pertaining to the simulation mode, the number of modules of each type is unlimited. The functions are also not restricted to the list described above. For example, a traffic module can function in parallel with multiple markets to guide self-driving EV participants to find optimal paths to carry passengers in conjunction with charge and discharge locations to maximize profit. Other modules that do not use traffic data to make decisions simply do not know about its existence.

A.3.4 Implementing TE in T-REX

TE systems can be setup in T-REX using a simple JSON configuration file. The T-REX runner assembles the modules as configured and launches them as independent processes on the assigned machine or machines at run time. Examples of configuration files can be found on the GitHub repository [64].

To launch a classical TC simulation, the following modules are required:

1. A non-participant powerflow module. The built-in implementation uses OpenDSS [100] and its Python API [101].

2. TC Market with a sub-module that generates prices based on the received powerflow data.

3. Participants containing load profiles, controllable devices, and price-reactive logic.

Figure A.3 shows a simplified version of the sequence flow diagram of the TC co-simulation implemented in T-REX. Due to the asynchronous nature of T-REX, many independent asynchronous functions and parallel loops have been omitted from the diagram, and only an approximation of the main flow path is shown.

In the same way, T-REX can also be configured to run agent-based economics (ACE) [102] simulations with minimal modifications from the TC configuration. In the example configurations, the only modifications are the removal of the parallel running powerflow module and swapping in the appropriate market module and agent logic submodules.

A.4 Double Auction Market Design for AI

A.4.1 Trading Mechanism

Price theory states that the price for any specific good or service is based on the balance of supply and demand. In a market-based TE approach, the role of the market is to efficiently facilitate the exchange of energy so that the price can appropriately and accurately reflect the balance of supply and demand at the time of exchange. ALEX adapts and adjusts an existing market design to fit three key considerations:

1. **Suitability for electricity grids with high penetration of DER and RES.** This means that, from a high level perspective, a market (or a collection of markets) must be able to effectively target localization and the intermittent nature of RES.

2. **Technical constraints and requirements of deployment:** Data acquisition, transportation, and cost must be minimized.

3. **Machine learning considerations for agents:** Related to the point above, ML will play an important role in trading and managing of energy resources in place of humans. For this reason, the market should be conducive to learning. One way to achieve this is to compose the market with a small set of explicit rules. The rules should provide a strong feedback signal, and they should be flexible enough to offer large action spaces.

With these considerations in mind, the final market is a modified form of double auctions [103][104]. The rules, explicitly implemented in the code, are described below:

1. It is assumed, for the time being, that the grid is an infinity bus and it can be interacted with through net billing. We therefore adapt retail electricity prices in Alberta, where buying energy from the grid costs \$0.1449/kWh, and selling earns \$0.069/kWh.

2. The local market has two energy pools: one for dispatchable sources, such as battery energy storage systems, and one for non-dispatchable sources, such as photovoltaics. This is intended to distinguish the source of energy, and to allow for the value of dispatchability to emerge.

3. Auctions settle for energy to be delivered during the one-round period from the end of the current round. However, the delivery period can be parametrically adjusted during run-time for future design explorations.

4. During the current round, participants submit bids and asks for energy to be delivered during or beyond the next delivery period.

5. A modified double auction system is used to settle trades: bids/asks are settled pairwise, with bids sorted from the highest to lowest, and asks in reverse to ensure pareto equality.

6. Bid/ask quantities can be partially settled.

7. A bid/ask quantity must be an integer multiple of 1 Wh. This is in consideration of future hardware integration, to allow direct use of the watt-pulse function of most smart-meters.

8. During the delivery period, if a seller is in short supply, it must financially compensate for the shortage at net metering prices. If batteries are available, the seller has the option to compensate by discharging its batteries, for all or part of the shortage during this period.

9. During the delivery period, if a buyer settled for more energy than used, the buyer must still pay the seller for the unused energy at the settlement price.

This market design strikes a compromise between a peer-to-peer market and a centralized market. By using pairwise settlements, a peer-to-peer like individualized value feedback can still be provided, while the simplicity and efficiency of a centralized market can be kept, especially for deployment in a small, localized region.

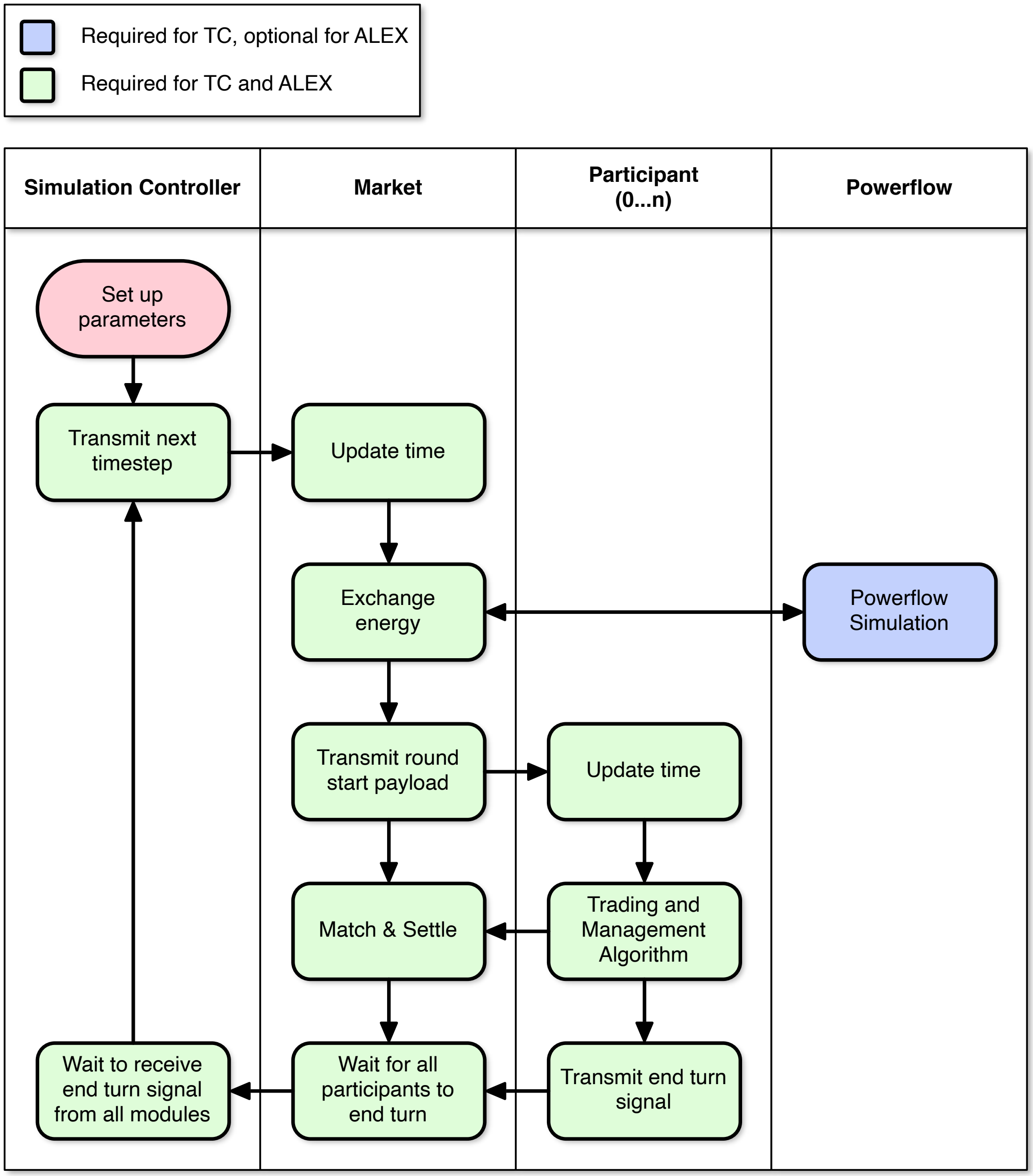

Figure A.3: Simplified swimlane diagram of TE schemes implemented in T-REX

Appendix B: Appendices for Chapter 3

B.1 Additional Information on Experiments

This appendix provides additional information on the experiments described in Section 3.3 and discussed in Section 3.4. It serves to further illustrate the performance of ALEX beyond the narrow focus on the discussed hypotheses.

Figures B.1 and B.2 depict the average daily community net load and average daily community state of charge (SoC), respectively, separated into the four seasons. We observe the same trends discussed in Section 3.4, under the influence of seasonal variance of load demand and photovoltaic power availability. The NoDERMS scenario provides the seasonal trend of net-load swing, which is most pronounced in Spring and Summer. While ALEX consistently reduces community net-load swing beyond the IndividualDERMS capabilities, its performance advantage is more pronounced as the seasonal net-load swing increases, due to it's capability to shift load within the community.

Figure B.3 depicts the average cumulative electricity bill for the three examined scenarios. Although the economic performance of ALEX is not directly relevant to the hypotheses posed in this article, the economic performance of LEM with respect to net-billing scenarios, such as IndividualDERMS, is often discussed in the LEM literature [67].

The economic welfare of a specific baseline depends on the grid sell price $p_{\text{grid,sell}}$ and grid buy price $p_{\text{grid,buy}}$ in a given jurisdiction, along with the accessible profitability gap. Typically, the difference between $p_{\text{grid,sell}}$ and $p_{\text{grid,buy}}$ comprises various fees or fee-like components. ALEX consistently outperforms IndividualDERMS in terms of economic welfare for the same setting, as long as a profitability gap exists. While assuming the existence of a profitability gap is not unrealistic, access to the full profitability gap remains a strong assumption [19]. However, performance across scenarios remains consistent for all other metrics, irrespective of the actual size of the profitability gap.

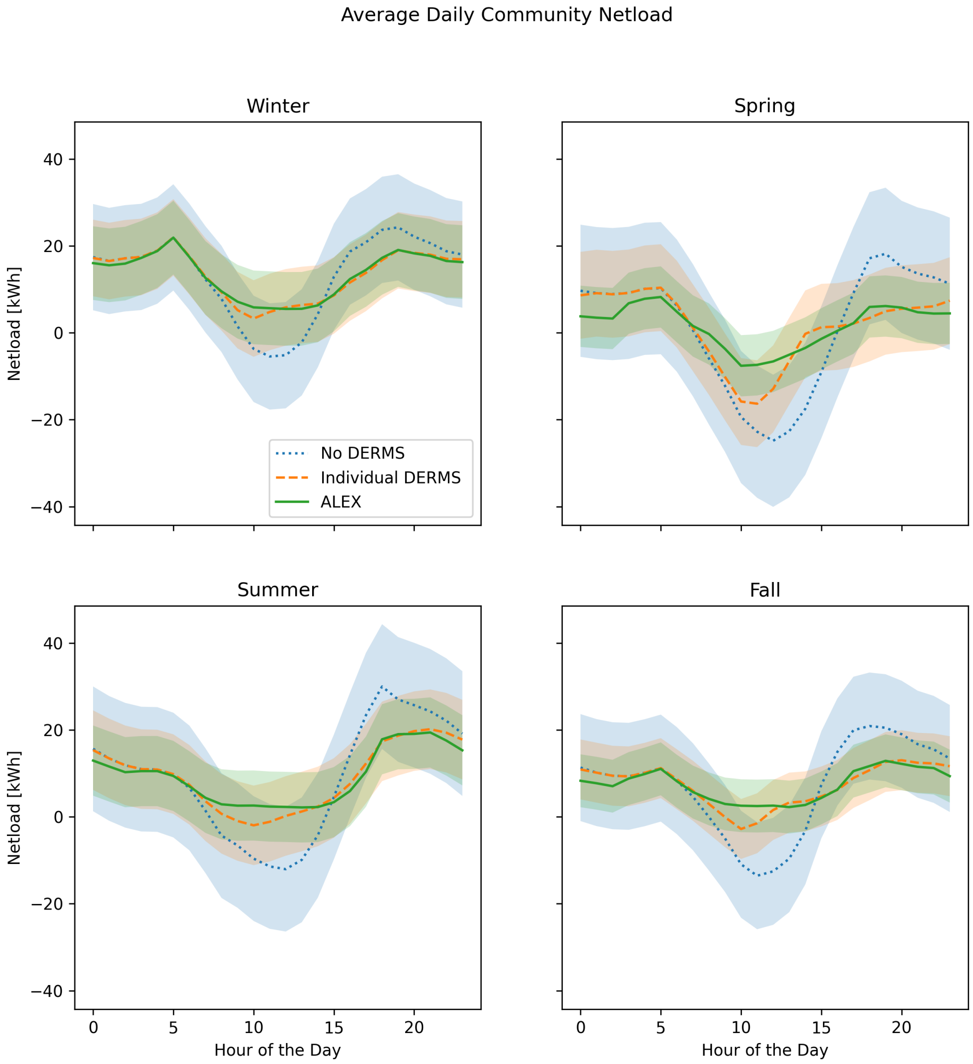

Figure B.1: Average daily net loads in kWh at hourly resolution for winter, spring, summer, and fall in a full simulation on the CityLearn 2022 data set [70] are presented for NoDERMS, IndividualDERMS, and ALEX scenarios. The figures display average values along with standard deviation bands.

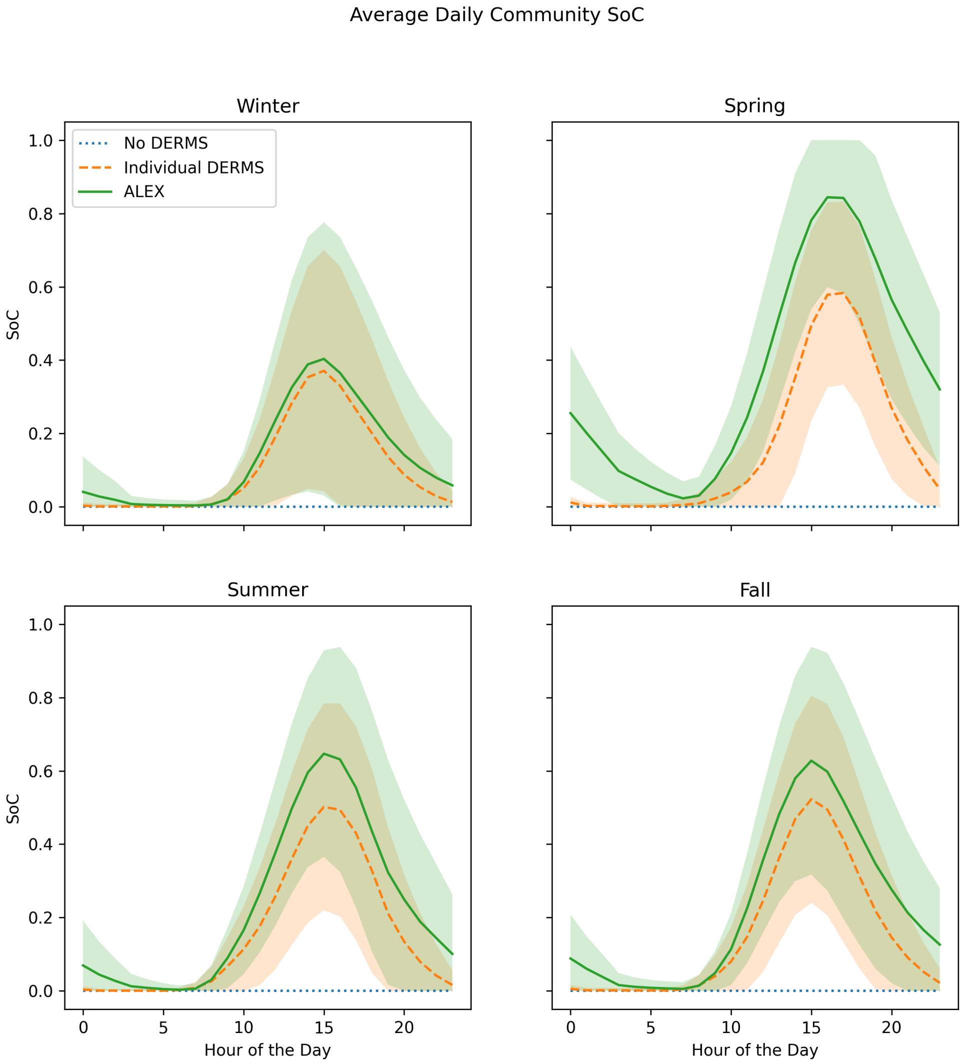

Figure B.2: Average daily SoCs at hourly resolutions for winter, spring, summer and fall of a full simulation on CityLearn 2022 data set [70] for NoDERMS, Individual-DERMS and ALEX scenarios. Shown are the average values as well as the standard deviation bands.

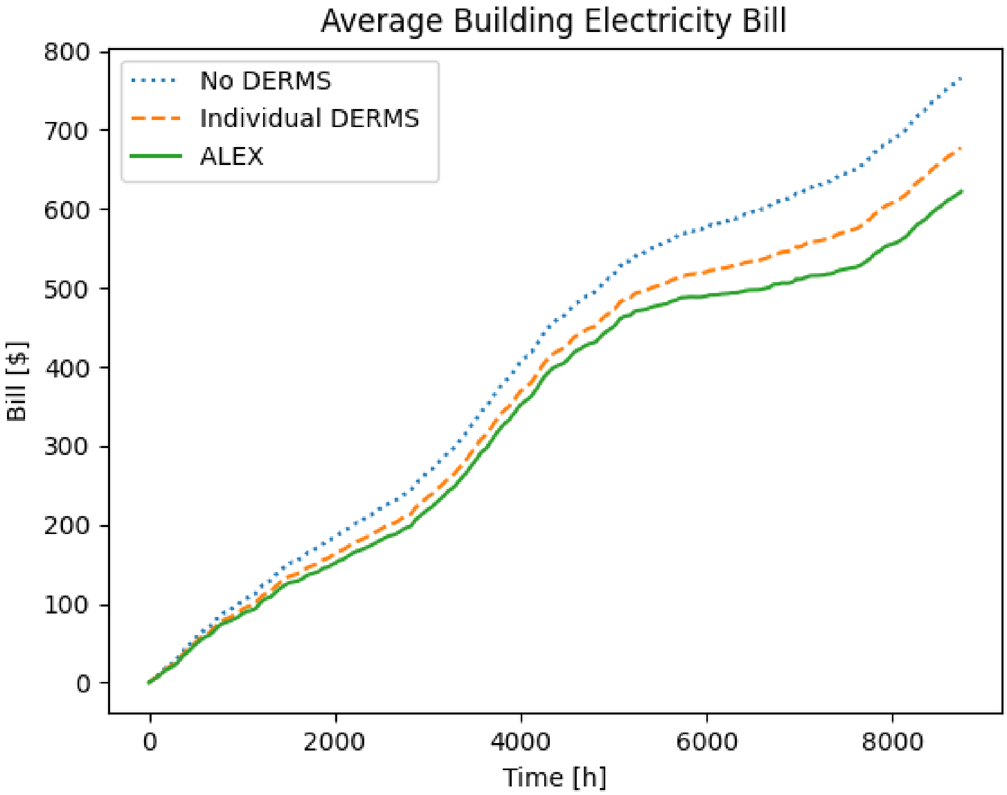

Figure B.3: Average cumulative building bill for a full simulation on CityLearn 2022 data set [70] for NoDERMS, IndividualDERMS and ALEX scenarios. In the depicted scenario, ALEX has access to the full profitability gap.

Appendix C: Appendices for Chapter 4

C.1 Hyperparameters

This appendix aims to enhance the reproducibility of the presented results by providing hyperparameters while also detailing the general approach taken in designing the DRL algorithm and testing the modifications.

The algorithm employed in this study is rooted in the publicly accessible Recurrent PPO implementation from Stable Baselines3 (SB3)[91]. The hyperparameter values that deviate from SB3's recurrent PPO default settings are as follows:

- The neural network architecture for both critic and actor consisted of 2 LSTM layers with 256 neurons each, followed by a 64-neuron head, along with a shared 64-neuron feature encoder.
- The actor's log standard deviation is initialized as -10 instead of the default 0.
- An exponentially decaying learning rate schedule is employed, reducing the learning rate by a factor of 0.69 every 1 million steps.
- The size of one mini-batch is set to 72, equivalent to one 3-day trajectory, based on SB3's recurrent PPO implementation for sample collection.
- The replay buffer stored 3672 transitions, equivalent to 9 days at 24 steps per day for 17 houses.
- The burn-in period for a single sample is set at 50% of the sample's length, or 36 steps.

The algorithm adaptations, design, and hyperparameter choices underwent testing across increasingly complex versions of the experiments discussed in the main body of this article until the performance detailed in the discussion section was achieved. The testing progression began with artificial load profiles, aiming to optimize net billing, then advanced to optimizing net billing on the City Learn dataset for a singular month, then the full year, and finally transitioned to the target application.

The advantage of this iterative process lies in the clearly defined optimal returns for the test scenarios. Recurrent PPO variants seem to vary across implementations, as the exact nature of making PPO recurrent is up to interpretation. We refer to Pleines et al. [105] for an investigation into the characteristics and sensitivities of recurrent PPO. Tests commenced with the default SB3 recurrent PPO in a shared experience replay setting [89], followed by the implementation of R2D2 [32], then state recalculation [34], and finally incorporating a learning rate schedule [33]. Each implementation underwent testing over a small range of hyperparameters for three runs each to ensure consistency, leading to the crystallization of the hyperparameter set used in this study.

The quality of a run was primarily evaluated based on its achieved return, supplemented by the investigation of various RL agent performance metrics. These metrics, inspired by those discussed in the SB3 documentation and Huang et al.'s insightful blog [106], encompassed explained variance, KL-divergence, and entropy loss curves. Even if an algorithm change did not directly impact the agent's average return, it was considered an improvement if, for example, it led to higher explained variance and thereby a stronger critic.

This iterative practice enabled the authors to initiate algorithm development in smaller, constrained versions of the final application, gradually scaling the difficulty of the experiments as the algorithm matured. Consequently, the algorithm utilized in this study is relatively basic and does not entail a vast array of modifications, focusing instead on targeted adaptations aimed at enabling the agents to construct a robust temporal state representation.