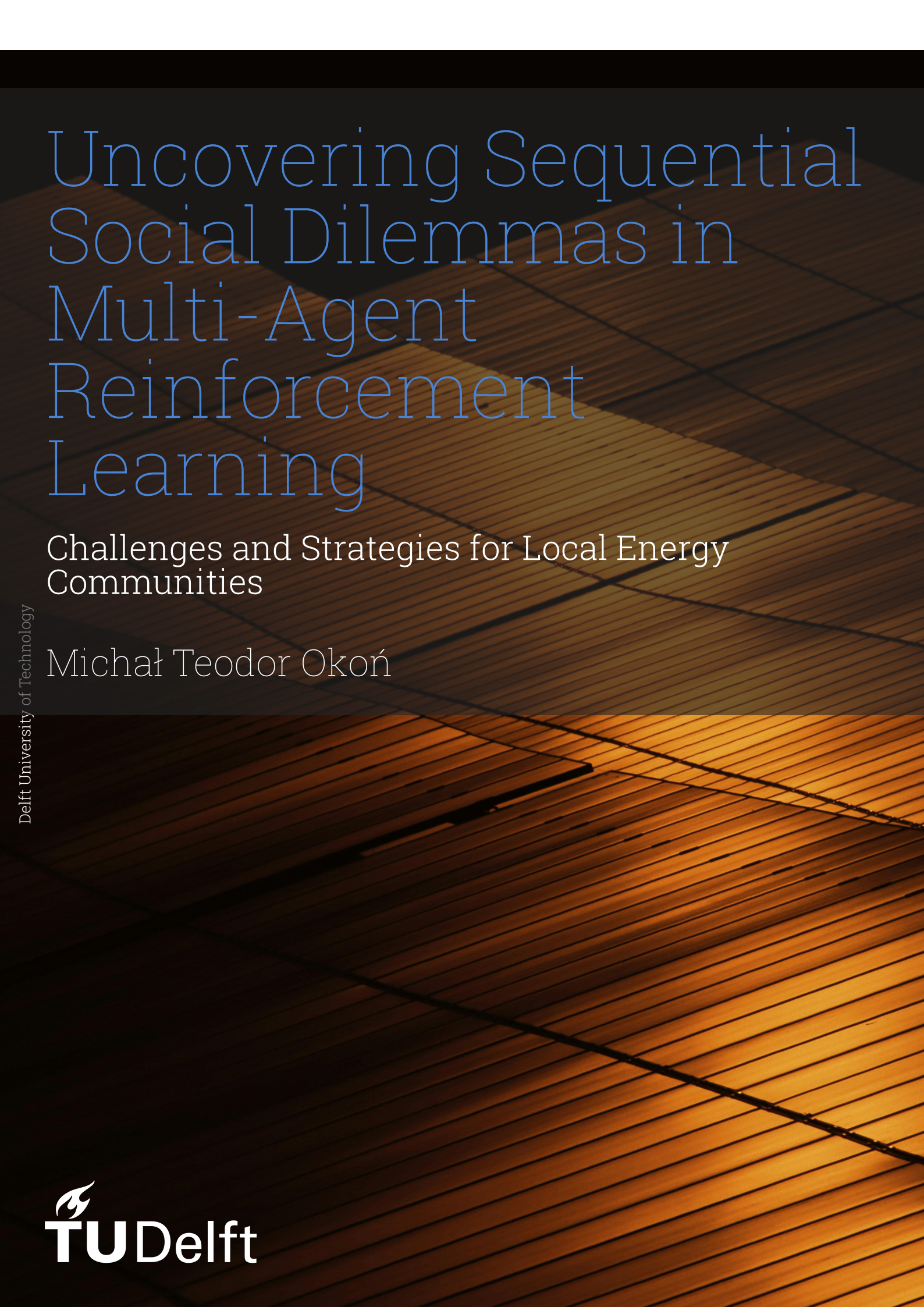
Uncovering Sequential Social Dilemmas in Multi-Agent Reinforcement Learning

Challenges and Strategies for Local Energy Communities

Michał Teodor Okoń

Delft University of Technology



TU Delft

Uncovering Sequential Social Dilemmas in Multi-Agent Reinforcement Learning

Challenges and Strategies for Local Energy Communities

by

Michał Teodor Okoń

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on 20 February 2025 at 13:00

Student number: 5056640

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

Summary

Local Energy Communities (LECs) have great potential to become one of the pivotal elements of the green energy transition. These decentralized energy systems offer participants a range of benefits, including increased sustainability, self-sufficiency, and reduction in both costs and CO2 emissions. With many different actors involved, various strategies have been implemented to effectively manage such communities, with Multi-agent Reinforcement Learning (MARL) standing out as a particularly promising approach. However, even with sophisticated techniques in place, management of LECs remains challenging as the community-wide objectives often conflict with the personal goals of the involved agents. These dynamics give rise to Sequential Social Dilemmas (SSDs)—scenarios where the choice to cooperate or defect no longer depends on a single action but unfolds over time. The temporal nature of SSDs, along with the inherent misalignment between personal and collective goals, makes it difficult for the RL algorithms to learn policies that optimize costs for both individual households and the whole community.

The main objective of this thesis is to understand how SSDs occur in LECs where MARL is used to guide the actions of individual agents. The first research question we pose asks how SSDs can be effectively detected and analyzed in a MARL-based LEC setting. The second question calls for concrete learning strategies that can be implemented to mitigate their negative effects and encourage cooperation among agents. To study these challenges, we develop a custom MARL environment where agents can interact through a shared energy pool and a local trading mechanism, creating conditions for us to uncover the existence of SSDs.

Several experiments are performed to systematically investigate these issues and measure the effectiveness of possible solutions. First, we focus on analyzing aspects of the environment where no cooperation is involved, showcasing how limited resources negatively impact the learning process. Next, we shift our focus to a more social setting where agents interact through a shared battery system. Using policy-matching techniques, we confirm that SSDs are indeed present and analyze how greed and fear factors influence agents to make choices that reduce community welfare. It is also noted that rescaling of the training data is an effective SSD mitigation technique leading agents to adopt more cooperative behaviors. In the third experiment, we propose another mitigation strategy by adding new components to the reward function to incentivize community-friendly battery use. This approach has proven to be successful as it results in more than a 23% increase in the social welfare of the community. Lastly, we verify how the proposed mitigation strategies perform in a more complex, realistic LEC environment where multiple, distinct households engage in trading and storage actions. The results align with our initial expectations - both rescaling and reward modifications boost households' cooperativeness and, in turn, improve the efficiency of our learning methods.

This thesis makes the following contributions:

1. We propose a new agent-centric approach to modeling LECs and develop an environment based on that.
2. We demonstrate that SSDs inherently occur in LECs, affecting the agents' abilities to learn behaviors beneficial to the whole community.
3. We propose two solutions for modifying learning procedures to better align community and individual incentives in LECs.

Finally, we suggest multiple directions that future research may follow, including the integration of community-oriented incentives into the trading mechanisms and utilizing more refined reward modeling with direct input from the agents.

Contents

1

Introduction

Local Energy Communities (LECs) have emerged as promising decentralized energy systems and an important pillar in the ongoing green energy transformation [1]. The main goal of such communities is to boost sustainability, self-sufficiency, and efficiency of local energy networks by allowing neighbors to exchange generated renewable energy. When combined with smart energy storage and demand-response management, the use of these local networks can lead to a great reduction not only in costs but also in CO2 emissions [2, 3]. Moreover, by providing democratic and equitable access to energy resources, these communities also exert a considerable social impact on their members [4]. In recent years, researchers have developed various models and techniques to optimize such communities. Among them, Reinforcement Learning has proven to be a particularly effective method, yielding outcomes that are highly reflective of real human incentives and adapting well to high levels of unpredictability [5, 6].

Managing energy communities, however, is not without its challenges as an interplay between social and technical dynamics can easily lead to suboptimal outcomes. In these settings, the interactions among households create a fertile ground for social dilemmas—situations in which individual rational choices lead to suboptimal outcomes for the group [7]. Such scenarios are usually modeled as games where the payoff of individual participants depends not only on their own actions but also on the actions of the whole collective. A well-known example of such a dilemma is the Prisoner's Dilemma [8] where each of the two players may choose to defect to earn a high reward. However, when both players take this action, a penalty is assigned to the two of them instead. Similarly, in the context of LECs, social dilemmas may arise when certain community members decide to exploit the shared energy pool for personal gain or some households are systematically excluded from the energy transactions, undermining the idea of energy justice.

In real-world cases like LECs, the concept of social dilemmas extends beyond single actions. In the context of Markov games, Leibo et al. [9] coined the term *Sequential Social Dilemma* (SSD) where the choice to cooperate or defect is no longer considered an atomic action but a policy that the agents employ. The researchers pointed out that in real-world settings, these choices are extended over time with cooperativeness reflected in the policies and not single actions which closely resembles the mechanisms behind actual social interactions.

Within a multi-agent reinforcement learning (MARL) framework, SSDs often reduce training efficiency and lead to suboptimal outcomes [9]. Despite the recent developments in the field of reinforcement learning, it is still challenging to account for the situations where agents engage in competitive interactions that still require varying degrees of collaboration over extended periods [10]. These dynamics complicate the decision-making process because each agent must adapt its strategy based on the changing strategies and actions of other agents. A range of different mechanisms for learning and adaptation is required to effectively handle the intricate aspects of such environments, ensuring that agents extend their learning beyond immediate returns to also support long-term group objectives. Some examples of such mechanisms include Contracting [11] where the agents' reward function is

modified through formal contracts and Gifting [12] where agents are equipped with the ability to gift rewards to others. However, both approaches have been evaluated only in highly abstract settings.

Despite theoretical progress in understanding SSDs, little attention has been paid to detecting and mitigating them in environments closely resembling real-world scenarios. This thesis aims to fill in this gap by investigating this problem from the perspective of Local Energy Communities. We argue that these communities present a compelling real-world setting where social dilemmas arise naturally and can be effectively studied.

This thesis examines the emergence and impact of Sequential Social Dilemmas in domains utilizing multi-agent reinforcement learning, with a primary focus on their effect on the learning process in the LEC environments. We also propose straightforward strategies to help mitigate these dilemmas and improve the training efficiency. Our research attempts to answer two critical questions:

1. How can Sequential Social Dilemmas be detected and evaluated within a multi-agent deep reinforcement learning framework applied to a Local Energy Community setting?
2. What reinforcement learning strategies and techniques can help overcome the challenges introduced by SSDs, improving the learning process and inter-agent coordination within a LEC environment?

To address these research questions, we have developed a custom reinforcement learning environment to accommodate LECs. This environment not only highlights the occurrence of SSDs in a reinforcement learning setting but also closely resembles real-world challenges often faced by these communities such as the need to balance between individual incentives and collective goals.

The contributions of this thesis are as follows:

1. We developed a novel RL-driven Python environment for investigating the emergence and mitigation of SSDs in the context of Local Energy Communities, incorporating energy storage and local energy trading.[1]
2. We demonstrated the existence of SSDs within these communities using our custom environment.
3. To enhance social welfare, we proposed two mitigation strategies- one involved rescaling the training data while the other altered the objective function to be more reflective of community goals.

The thesis is structured in the following way. In Chapter 2, we provide the necessary background on Reinforcement Learning, Social Dilemmas, and Local Energy Communities, and explain how these concepts intersect to form the foundation of our research. Chapter 3 details our approach to modeling the LEC from both theoretical and reinforcement learning perspectives. Building on this model, the relevant experiments are conducted in Chapter 4. Finally, Chapter 5 summarizes our findings.

[1] GitHub repository with the code: `https://github.com/MichalOkon/marl-lec`

2

Background

2.1. Reinforcement Learning

2.1.1. Single-Agent Reinforcement Learning

Reinforcement Learning (RL) is a part of machine learning where agents learn from their own actions. Instead of being directly given the correct answers (like in the supervised learning), RL agents gain knowledge about the environment by exploring, taking actions, and earning rewards (or penalties) depending on their decisions [13]. Internally, RL is based on Markov Decision Processes (MDPs) which provide a mathematical framework for modeling situations where outcomes are somewhat uncertain yet still influenced by the agent's actions.

MDPs are structured in the following way. Let M be a Markov Decision Process. Then M can be defined as a tuple in the following way:

$$M = (S, A, P, R)$$

where:

- S: All possible states the system can be in.
- A: All possible actions the agent can take.
- P: The transition probability function $P : S \times A \times S \to [0, 1]$, giving the odds of moving to state s' from state s after taking action a.
- R The reward function, $R : S \times A \to \mathbb{R}$, which informs us how beneficial taking a is in state s is.

The goal of the agent is to find a policy π (a function that maps states to agent's actions) maximizing its cumulative reward over time. The cumulative reward function takes the following form:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Here, R_{t+k+1} is the reward received after k transitions from time t and γ is the discount factor, a real number in the interval $[0, 1]$, which purpose is to make future rewards less valuable than immediate ones.

One of the tasks of an agent in such an environment is to determine which action yields the highest reward in a given state. To that end, the agents uses two key value functions:

- State-Value Function $V^\pi(s)$: This tells us how rewarding it is to start in state s and follow the policy π:

$$V^\pi(s) = \mathbb{E}\left[G_t \mid S_t = s, \pi\right]$$

- Action-Value Function $Q^{\pi}(s, a)$: This measures how beneficial it is to take action a in state s and then proceed with policy π:

$$Q^{\pi}(s, a) = \mathbb{E}\left[G_t \mid S_t = s, A_t = a, \pi\right]$$

2.1.2. Multi-Agent Reinforcement Learning

Multi-Agent Reinforcement Learning (MARL) expands RL by introducing multiple agents that interact with each other and the environment. In MARL, agents must adapt not only to the environment, which can be a very challenging problem by itself but also to each other's actions, adding more complexity to such a system. MARL is a robust framework for tasks like driving, robotic swarms, and economic modeling [14, 15, 16].

The foundation of MARL is Markov Games (also known as Stochastic Games), which generalize Markov Decision Processes to multi-agent settings [17]. A Markov Game for N agents can be defined by the tuple:

$$G = (S, A_1, \ldots, A_N, P, R_1, \ldots, R_N)$$

where:

- S: Set of all possible states of the environment.
- A_i: Set of actions available to agent i (where $i \in \{1, \ldots, N\}$).
- P: Transition probability function, $P : S \times A_1 \times \cdots \times A_N \times S \to [0, 1]$, which defines the likelihood of moving to state s' from state s after the agents take actions $(a_1, \ldots, a_N)$.
- R_i: The reward function for agent i, which depends on the state and the actions taken by all the agents.

In MARL, each agent's goal is to find a policy $\pi_i : S \to A_i$ that maximizes its own cumulative reward:

$$G_t^{(i)} = R_{t+1}^{(i)} + \gamma R_{t+2}^{(i)} + \gamma^2 R_{t+3}^{(i)} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}^{(i)}$$

where $R_{t+k+1}^{(i)}$ is the reward received by agent i after k transitions, and γ is the discount factor.

As in single-agent RL, MARL often relies on value functions, but, in this case, these functions depend on the actions and policies of multiple agents:

- State-Value Function $V^{\pi_1,\ldots,\pi_N}(s)$: The expected cumulative reward when all agents follow the joint policy $(\pi_1, \ldots, \pi_N)$ starting from state s:

$$V^{\pi_1,\ldots,\pi_N}(s) = \mathbb{E}\left[G_t^{(i)} \mid S_t = s, \pi_1, \ldots, \pi_N\right]$$

- Action-Value Function $Q^{\pi_1,\ldots,\pi_N}(s, a_1, \ldots, a_N)$: The expected cumulative reward when starting from state s, taking the actions $(a_1, \ldots, a_N)$, and then following the joint policy $(\pi_1, \ldots, \pi_N)$ thereafter:

$$Q^{\pi_1,\ldots,\pi_N}(s, a_1, \ldots, a_N) = \mathbb{E}\left[G_t^{(i)} \mid S_t = s, A_t^{(1)} = a_1, \ldots, A_t^{(N)} = a_N, \pi_1, \ldots, \pi_N\right]$$

MARL presents unique challenges that differ from single-agent environments. A major issue is non-stationarity. As each agent is learning and adapting simultaneously, the environment keeps changing from the perspective of a single agent, which complicates the learning process [18]. Additionally, partial observability often obstructs each agent's view of the global state, forcing them to make decisions based on incomplete information [10]. With the growing number of agents, scalability becomes an issue. The complexity of the state and action space tends to increase exponentially, making it difficult to compute optimal policies [16]. Lastly, in cooperative settings, agents need to work together, which often makes it necessary to share information or align strategy [19]. These are just a few examples of the hurdles that make multi-agent environments much more difficult to work with than the single-agent variations.

2.1.3. Proximal Policy Approximation

Proximal Policy Optimization (PPO) is one of the most widely-used RL algorithms and is frequently used in the LEC training environments [20, 21]. Introduced by Schulman et al. [22], this policy gradient method was designed to increase the stability and efficiency of training by avoiding making too large, destabilizing changes to the employed policy while still improving its performance. One can think of it as a middle ground between the simplicity of standard policy gradient methods and the more complex techniques like Trust Region Policy Optimization (TRPO) [23].

The key innovation in PPO lies in its objective function, which limits how much a policy can change during each update. This helps stabilize the learning process while avoiding drastic, destabilizing shifts that other methods might allow. The algorithm achieves this by clipping the ratio of probabilities between the new and old policies, keeping updates within a safe range.

PPO works in the following way: Let π_θ be the policy parameterized by θ, and $L(\theta)$ represent the PPO objective function, which is defined as:

$$L^{CLIP}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right]$$

where:

- $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the probability ratio between the new policy π_θ and the old policy $\pi_{\theta_{\text{old}}}$.
- $\hat{A}_t$ is the advantage estimate at time step t, which measures how much better an action is compared to the expected action under the current policy.
- ϵ is a hyperparameter that sets the clipping threshold, typically a small value like 0.1 or 0.2.

By clipping the probability ratio $r_t(\theta)$ within the range $[1-\epsilon, 1+\epsilon]$, PPO penalizes updates that would otherwise push the policy far from its previous version. This ensures that the learning process remains stable but allows enough flexibility for exploration and continuous improvement. It is also very sample-efficient as it can re-use data from a single trajectory multiple times making it suitable for environments where data collection is costly.

PPO has also proven effective in MARL setups. By employing clipping, PPO manages the stability of policies across agents and allows them to learn effectively without destabilizing each other. In cooperative settings, PPO is often used with techniques like Centralized Training with Decentralized Execution (CTDE), where agents learn shared representations during training but act independently in execution. Examples of multi-agent scenarios where PPO was successfully employed include autonomous vehicle coordination, robotic swarms, and competitive games [24].

2.2. Social Dilemmas

Social dilemmas, or collective action problems, are situations where individual decision-makers face a conflict between personal interests and collective well-being [25, 26]. Although individuals would benefit more from cooperating, they often do not due to conflicting interests. This leads to various problems when actors pursue personal short-term objectives instead of acting in the group's best interest. Social dilemmas have been studied from multiple disciplinary perspectives, including economics [26], psychology [27], and political science [28].

Aside from the Prisoner's Dilemma, another relevant example of a social dilemma is the Tragedy of the Commons [29]. In this scenario, individuals benefit from maximizing their own usage of a shared resource, but overuse leads to depletion, harming everyone. Other notable examples of games include Stag Hunt [30] and Game of Chicken [31]. In the Stag Hunt, players decide to either hunt for a stag together gaining the highest reward, or go after a hare which is a safer option but awards less reward. Meanwhile, Game of Chicken is an example of a game where both players face a disaster if none of them decide to yield and accept a lower reward.

In Reinforcement Learning, Macy et al. [32] formalized the concept of a Social Dilemma as a mixed-motive two-player game where players choose to either cooperate or defect. These choices lead to four different payoffs: R (Reward) for mutual cooperation, P (Punishment) for mutual defection, S (Sucker) when one cooperates and the other defects, and T (Temptation) when the situation is reversed. The

outcomes and their relation to defection/cooperation behaviors are illustrated in picture 2.1. For a social dilemma to occur, the following criteria must be met:

1. $R > P$: Mutual cooperation yields a higher reward than mutual defection.
2. $R > S$: Mutual cooperation is more beneficial than being cheated on.
3. $2R > T + S$: The sum of rewards for mutual cooperation is higher than the sum of rewards from unilateral cooperation and defection.

Additionally, one or both of the following conditions must be true:

1. $T > R$: Unilateral defection is more rewarding than mutual cooperation, tempting players to cheat.
2. $P > S$: Mutual defection is more rewarding than unilateral cooperation, leading players to defect out of fear.

Depending on which conditions are met, there are three possible outcomes:

- If only $T > R$ holds, it is equivalent to a Game of Chicken.
- If only $P > S$ holds, it is equivalent to a Stag Hunt game.
- If both $T > R$ and $P > S$ hold, it is equivalent to the Prisoner's Dilemma.

All the games discussed so far are usually modeled as one-shot games, where players make a one-off decision, emphasizing the immediate strategic nature of their actions. However, such situations are rare in real-world applications. To address this, Leibo et al. [9] extended the definition by introducing a temporal element, creating the concept of Sequential Social Dilemmas. This approach replaces single actions with long-term policies executed by the agents, making the model more applicable to real-world scenarios where problem-solving techniques like Reinforcement Learning can be utilized.

	Cooperate	**Defect**
Cooperate	R, R	S, T
Defect	T, S	P, P

Figure 2.1: Matrix-based representation of a social dilemma in a one-shot setting.

Plenty of scientific efforts have been made to improve the efficiency of Reinforcement Learning learning models in environments where Social Dilemmas are prevalent. Lerer et al. [33] proposed a solution in which the agents learn to evaluate the cooperativeness level of their opponent's actions and reciprocate the sentiment of their adversary with their own actions. Meanwhile, Yu et al. [34] examined the domain of Spatial Social Dilemmas where agents form a network of interacting entities. To make the agents more aligned with human values, the researchers included emotion-related mechanisms in their system which impacted the learning behaviors of the agents. On the other hand, Anastassacos et al. [35] explored the dynamics of cooperation and reputation within multi-agent systems where social dilemmas can be present, demonstrating that reputation mechanisms can lead to stable cooperation under specific conditions. Finally, Lupu et al. [12] investigated how endowing the agents with the ability to gift rewards to others affects the learning setting, while Haupt et al.[11] looked into formal contracting as a way of enhancing the learning process.

2.3. Local Energy Communities

Local Energy Community (LEC), often simply referred to as an energy community, is defined as a group of local energy consumers and producers collectively engaged in the generation, consumption, storage, and management of renewable energy [36]. The main goal of such communities is to boost energy efficiency, sustainability, and security. Aside from that, energy communities empower consumers giving them more means to manage their own energy systems, democratizing the energy exchange, and providing economic benefits [4, 2].

The following section details the social impact as well as challenges faced in such communities. As LECs constitute a very broad topic, the exact implementation details and mechanisms employed can vary widely in the literature. To provide a solid technical foundation for the readers, we also look into the most commonly researched and utilized aspects of energy communities which include energy generation, storage, and trading.

2.3.1. Social Impact

Due to the strong social component of this thesis, we start by discussing the impact social dynamics have on energy communities. According to Bielig et al. [4], energy communities play a crucial role in creating a social impact for their members by delivering economic, ecological, and social benefits. This impact can be grouped into four categories: energy justice, energy democracy, community empowerment, and social capital.

Energy justice builds on social justice by ensuring that both the benefits and burdens of energy systems are fairly shared among all members of society. This concept encompasses terms like distributional justice (fair sharing of benefits and burdens), procedural justice (ensuring that the decisions are made in fair and inclusive ways), and recognitional justice (involving marginalized groups) [37]. Energy democracy calls for more democratic control over energy systems through active participation of community members, as well as equal access to the decision-making process and shared ownership of the energy resources [38, 39],. By community empowerment, we mean improvement of local capacities by providing access to material, social, and knowledge resources [40, 41]. Finally, social capital refers to the strength of social networks and building mutual trust within a community allowing for effective collaboration and the overall success of such energy initiatives [42].

2.3.2. Challenges

Designing and operating a successful LEC is a complex challenge and a lot of scientific efforts are being invested into ensuring the efficiency and profitability of such communities. Among the most researched topics is cost optimization [2]. That encompasses minimization and economic analysis of costs related to investments [43], operations [44], energy [45, 46], greenhouse gas emissions [47], and battery costs [48]. Other studies focus on areas such as reduction of environmental impact [49], increasing energy independence [50], or load management [51]. Additionally, some research also explores the comfort needs of LEC members [52]. Single papers rarely address a single object, often focusing on many interconnected goals.

2.3.3. Energy Generation

Renewable energy technologies allow individual households to make savings, reduce their air pollution levels, and improve energy safety by supplying them with means to generate their own energy. There are many different sources of renewable energy that utilize means such as wind, hydro, an geothermal energy [2]. While there exist many different means for houses to generate energy, Photovoltaic (PV) panels constitute one of the most popular choices and the generation method we focus on in this thesis.

In simulations and optimization problems, PV panels can be modeled using a variety of approaches. One of the more popular methods is to employ a pre-existing simulation tool that generates energy production data based on the characteristics of the PV system and conditions of the environment[53, 54]. Alternatively, an analytical approach can be used where the energy output is calculated based on technical factors specific to the system such as the number of photovoltaic cells, module efficiency, and the panel surface area [54] with some methods also accounting for temperature [55] for greater accuracy. Lastly, the data might be gathered directly from real-life energy communities which provides very precise datasets [56].

2.3.4. Energy Storage

At the level of individual households, peak energy generation rarely coincides with periods of the highest energy demand often leading to wasted renewable energy [57]. Battery energy storage systems (BESS) address this issue by allowing households to store energy on either individual or community level [58]. This results in increased energy flexibility, cost reductions, and improved sustainability. Installed on an individual level, batteries allow for greater control over renewable consumption, enable load shifting,

and strengthen energy independence. Communal battery systems, also called Community Energy Storage (CES), offer perks such as cost-sharing, increased self-consumption of renewable energy, and peak demand shaving [59, 60].

From a technical viewpoint, the most commonly used kinds of batteries include lithium-ion batteries, known for their high energy density and efficiency, and lead-acid batteries, which are cheaper to install but less durable. Flow and sodium-sulfur batteries are also gaining popularity due to their scalability and suitability for large-scale communal storage needs [61].

Similar to the PV output modeling, several different battery models are employed in the literature [2]. Among these, an ideal model assumes no loss during charging and discharging [62]. The most widely used approach incorporates a simple charging/discharging efficiency model where a fraction of energy is lost during these actions [63]. Additionally, some papers account for self-discharge reflecting gradual energy loss over time [64] as well as effects related to degradation of batteries [65].

2.3.5. Energy Trading

Energy trading is widely considered one of the fundamental components of LECs. It allows the community participants to exchange energy reducing the reliance on the external grids [66]. Other benefits include a reduction of peak-to-average ratio and an increase in participant's welfare by allowing them to trade energy at more profitable prices [67].

Local Energy Markets (LEMs) can be divided into three categories: Full P2P Markets, Community-Based Markets, and Hybrid P2P Markets [68]. Full P2P Markets are defined as collectives where each member can freely exchange its energy through direct transactions with other members [69, 70] often utilizing technologies such as blockchain [71]. By contrast, in Community-Based Markets, negotiations take place through a third-party supervisor playing the role of a mediator and negotiator between involved agents [72, 73]. To facilitate the negotiations, auction-based mechanisms are often employed [73, 74] Mixing these two approaches, Hybrid P2P Markets operate by clustering energy users into communities. Usually, in this sort of market, network participants can interact with the users inside their community but the trade between communities takes place through intermediaries [75, 76].

2.3.6. RL in LECs

Advancements in renewable energy technology and the growth in control households have over their energy management have made it increasingly difficult to solve some of the LECs-specific problems. Traditional optimization methods such as Linear Programming [77, 78], or Mixed-Integer Linear Programming [79, 80] often fail to properly account for the inherent unpredictability of such systems. With the growing quantities of data at our disposal, we can face these challenges with more data-driven approaches such as Reinforcement Learning [6]. Problems most frequently targeted by RL-based solutions can be grouped into the ones relating to energy management (also referred to as dispatch problems), and focusing on discovering optimal energy trading strategies in local networks [81]. Many papers do not limit their study to a single problem, often introducing both dispatch and trading to their research [82]. Due to the field's diverse nature, various RL algorithms are employed for LEC optimization problems [81].

Energy management encompasses actions agents take to satisfy the energy demand while minimizing energy costs [83]. This includes the procurement of electricity, heat, and cooling by utilizing various energy sources. [84, 85]. Energy storage, on both local and community levels, is very often considered in such settings [86]. While out of the scope of this research, on a household level, RL can also be employed to manage building utilities such as air-conditioning, heating, and air ventilation[87, 88]. Demand response, understood as job scheduling for household appliances, is another problem considered by existing studies [89].

In environments that facilitate P2P trading, RL often plays a role in defining the trading policy. Depending on the employed trading mechanisms, the resulting policies may either control the quantities of energy traded with neighbors [90, 91] or additionally determine the dynamic pricing of the exchanged energy [92].To provide pricing flexibility, many environments employ double-sided auctions as a way for the agents to exchange energy [93]. In these cases, the RL algorithms aim at learning the optimal bidding strategy [94, 82].

3

New LEC Environment for Demonstrating SSDs

3.1. Our Solution

In the existing literature, Local Energy Communities have been modeled in a myriad of ways. Due to the multifaceted character of the field, these solutions differ in scope, scale, and available functionalities [3, 83, 95]. Depending on the use case, the focus shifts to various aspects of such systems, from long-term numerical optimization to real-time simulations and agent-based models. Due to the specific nature of this thesis, we decided to create a new environment where households are modeled as independent RL agents, whose every trading and storage-related action is driven by a trained policy.

As highlighted earlier, the primary aim of this thesis is to investigate Local Energy Communities in the context of social dilemma formation. Therefore, the environment for this purpose must strike a balance between simplicity and realism. It should be simple enough to be generalizable to other problems and remain tractable while maintaining a high level of similarity to real-life local energy communities to ensure it is well-grounded in reality. The system should also be modular to allow examination of a range of scenarios where social dilemmas arise, as discussed in subsection 3.2. From a practical perspective, the environment needs to align with existing Reinforcement Learning frameworks, which are predominantly Python-based. The limited number of papers with open-source codebases further complicated the task of finding a suitable environment among the pre-existing ones. For these reasons, a new environment has been developed to satisfy all the aforementioned requirements. The following section outlines its implementation.

The main building blocks of the environment are households which act as independent agents with the objective of maximizing private profits. A general overview of the environment can be seen in Figure 3.1. On the individual level, households may be equipped with PV panels and energy storage systems. However, households vary in terms of available utilities and their power, that is, some households may be equipped with batteries of different capacities and PV panels with varying efficiency, while others may lack these assets altogether. This inequality reflects many real-world scenarios where households are rarely equal regarding the range of renewable assets they possess. Houses can manage the energy in multiple ways with their own energy needs as a top priority. The energy loads are fixed meaning there is no way of postponing or shifting them. Moreover, they are specific to each household with some of them having higher demands than others.

3.1.1. Local Energy Exchange

Aside from the energy exchange with the external grid, households can engage in trade among themselves by participating in the Local Energy Market. To make the transaction profitable for both parties, a midpoint price between the retail import and export prices is used to make the transaction profitable for all parties involved. On a decision-making level, households can determine the quantity of energy they are willing to trade but do not choose who to trade with. If more than two households participate in a

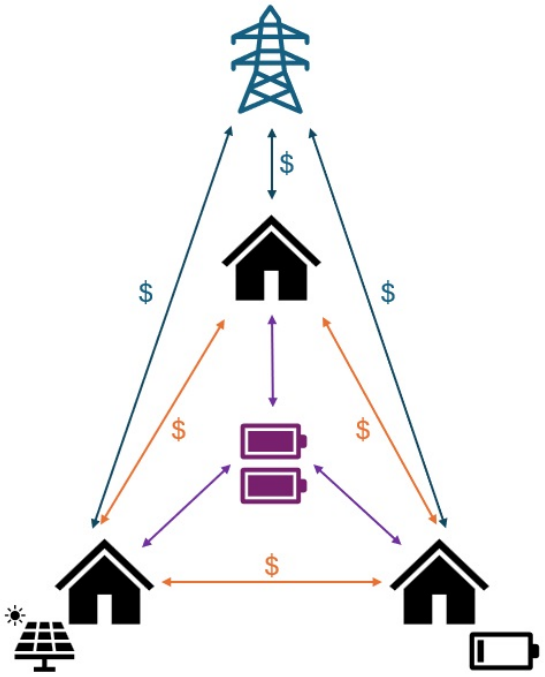


Figure 3.1: Overview of the complete environment. Households may include PV panels and local energy storage, with the option of a shared battery (highlighted in purple). Energy exchanges can occur between households (highlighted in orange) and with the retailer (highlighted in blue).

trade, the costs, profits, and energy volumes are split proportionally to the level of involvement among all the trade participants. Below we present energy exchange equations for a local energy market under different supply and demand scenarios.

The variable definitions are as follows:

- D_i: Demand of buyer i.
- S_j: Supply of seller j.
- P: Fixed market price.
- Q_i: Energy allocated to buyer i.
- Q_j: Energy allocated from seller j.

First, we start by defining Total Demand and Supply:

$$T_D = \sum_i D_i$$
$$T_S = \sum_j S_j$$

If Total Demand is equal to Total Supply, then simply:

$$\begin{aligned} Q_i &= D_i, \\ Q_j &= S_j, \\ \text{Cost}_i &= Q_i \cdot P = D_i \cdot P, \\ \text{Revenue}_j &= Q_j \cdot P = S_j \cdot P. \end{aligned}$$

If Total Supply exceeds Total Demand, then:

$$\begin{aligned} Q_i &= D_i, \\ Q_j &= S_j \cdot \frac{T_D}{T_S}, \\ \text{Cost}_i &= Q_i \cdot P = D_i \cdot P, \\ \text{Revenue}_j &= Q_j \cdot P = S_j \cdot \frac{T_D}{T_S} \cdot P. \end{aligned}$$

Otherwise, Total Demand exceeds Total Supply, meaning that:

$$\begin{aligned} Q_j &= S_j, \\ Q_i &= D_i \cdot \frac{T_S}{T_D}, \\ \text{Cost}_i &= Q_i \cdot P = D_i \cdot \frac{T_S}{T_D} \cdot P, \\ \text{Revenue}_j &= Q_j \cdot P = S_j \cdot P. \end{aligned}$$

3.1.2. Environment Setup

In our work, we consider a dispatch problem with the goal of managing the storage and trading systems efficiently in order to minimize energy costs. Generally speaking, the environment is model as a general-sum Markov game where agents operate simultaneously. Actions within the environment are executed in the following order:

1. Local energy market is processed. Energy and financial transfers are calculated based on the energy quantities agents are willing to buy or sell.
2. Agents interact with the community battery, deciding whether to charge or discharge it. The order in which the battery is accessed is randomized to ensure fairness.
3. Each agent takes actions relevant to its local environment:
   (a) Storage actions are executed. Energy is added to or withdrawn from the energy pool available to a household.
   (b) Depending on the energy balance, energy is either exported to or imported from the retailer.

At the end of each timestep, the household's energy balance is computed based on energy demand, storage interactions, and locally exchanged energy. Any surplus or deficit of energy is resolved through the retailer so that there is no unused energy or unmet energy demand at the end of each step. The energy balance is calculated based on the following equation:

$$E_{\text{balance}} = E_{\text{gen}} + E^r_{\text{im}} + +E^l_{\text{im}} + E^l_{\text{dis}} + E^s_{\text{dis}} - E_{\text{load}} - E^r_{\text{ex}} - E^l_{\text{ex}} - E^l_{\text{ch}} - E^s_{\text{ch}}$$

where:

- E_{gen}: Energy generated from PV panels.
- E^r_{im}: Energy imported from the retailer.
- E^l_{im}: Energy imported from local sources (other households).
- E^l_{dis}: Energy discharged from local storage.
- E^s_{dis}: Energy discharged from shared storage.
- E_{load}: Energy required to meet the household's demand.
- E^r_{ex}: Energy exported to the retailer.
- E^l_{ex}: Energy exported to local sources (other households).
- E^l_{ch}: Energy charged into local storage.

- E^s_{ch}: Energy charged into shared storage.

The load and generation data come from real-life datasets. At the end of each timestep, through exchange with the retailer, the system ensures:

$$E_{\text{balance}} = 0$$

If a household has PV panels installed, the generated energy is calculated as:

$$E_{\text{gen}} = \eta_{\text{pv}} \cdot E_{\text{total}}$$

where:

- E_{gen}: Energy generated in a 15-minute window.
- η_{pv}: PV coefficient, representing the fraction of the total energy generated from photovoltaic sources.
- E_{total}: Total energy generated in a 15-minute window.

Batteries have limited efficiency. For simplicity, we assume that this efficiency is fixed regardless of the current state of charge of the batteries. Therefore, the amount of energy that batteries are effectively charged with is calculated as:

$$E_{\text{stored}} = \eta_{\text{ch}} \cdot E_{\text{ch}}$$

where:

- E_{stored}: Actual energy stored in the battery after charging.
- η_{ch}: Charging efficiency, representing the fraction of input energy that is successfully stored.
- E_{ch}: Energy input used for charging the storage system in a 15-minute window.

In all the other cases, we assume that the losses incurred during energy transfer are negligible and we do not account for them.

On the implementation level, the environment is designed using the Ray RLlib framework which serves the purpose of a training library. The time series data has been divided into 15-minute time windows each representing a discrete step in the environment. During each step, agents independently observe the environment. Their knowledge is limited to the information specific to the respective household. The observation set is defined as:

$$O = \{P_i, P_e, L, L_n, E_{hh}, S_l, S_s, \mathbf{G}^f, \mathbf{P}^f\}$$

where:

- P_i and P_e: Current retailer energy import (P_i) and export (P_e) prices.
- L and L_n: Current energy load (L) and the energy load forecast for the next 15 minutes (L_n).
- E_{hh}: Energy exchanged with other households.
- S_l: Current state of charge of the local battery (if available).
- S_s: Current state of charge of the shared battery (if available).
- $\mathbf{G}^f = [G^f_1, G^f_2, G^f_3, G^f_4]$: A 4-hour forecast of energy generation (if PV panels are present), with 1-hour gaps between values.
- $\mathbf{P}^f = [P^f_1, P^f_2, P^f_3, P^f_4]$: A 4-hour forecast of retailer import prices, with 1-hour gaps between values.

To reduce the size of the observation space and reflect the uncertainty of the environment, forecasts are aggregated into 1-hour intervals. This results in a more concise representation with each forecast consisting of four averaged values.

Having this information, each household makes a set of actions. The action space encompasses storage and trading-related decisions. For both common and individual batteries, agents decide whether to charge or discharge the battery and what quantity of energy is to be used represented as a fraction of the battery's power. In the case of trading, the households choose whether and what type of trading action they want to make, as well as what amount of energy to trade.

The full action space takes the following form:

$$A = \{a_l, a_s, a_t, q_l, q_s, q_t\}$$

where:

- a_l and a_s: Categorical variables determining whether the local and shared batteries should be charged, discharged, or neither.
- a_t: A categorical variable determining whether the agent buys, sells, or does nothing in the local energy market.
- q_l and q_s: Fraction of the battery power to be used to either charge or discharge the local or shared batteries.
- q_t: Quantity of energy to be bought or sold in the local energy market.

The reward function focuses on minimizing the energy costs and is defined as:

$$r_t = \sum_t \left(P^r_{\text{im}} \cdot E^{\text{retail}}_{\text{im},t} + P^{\text{local}} \cdot \left(E^l_{\text{im},t} - E^l_{\text{ex},t} \right) \right)$$

where

- $\text{Cost}_{\text{total}}$: Total energy cost over the optimization period.
- t: Time index (15-minute timesteps).
- $P^{\text{retail}}_{\text{im}}$: Price per unit of energy imported from the retailer.
- $E^r_{\text{im},t}$: Energy imported from the retailer at time t.
- P^{local}: Midpoint price per unit of energy traded locally in the community.
- $E^l_{\text{im},t}$: Energy imported from local trades at time t.
- $E^l_{\text{ex},t}$: Energy exported through local trades at time t.

Notice that we do not include energy exported to the retailer in our equation. Initial experiments suggested that excluding this factor from the final reward greatly improved the training results as it encouraged the agents to focus on strategic actions such as optimizing energy storage rather than chasing immediate rewards.

3.2. Social Dilemmas in Local Energy Communities

As explained earlier, the primary focus of local energy communities is to increase the efficiency of local energy grids by enabling energy sharing among neighbors. However, this environment provides many opportunities for the exploitation of cooperative households by more self-centered ones. This section gives two examples of how such dynamics may come to life in our environment. In both cases outlined below, the energy pricing follows a Time-of-Use policy for imports from the retailer and fixed pricing for exported energy. Export prices are significantly lower than import prices, meaning that houses are encouraged to use the energy they produce rather than sell it, which is closely aligned with real-world scenarios. Additionally, each household needs to satisfy a certain energy load at any given time window using available energy sources.

3.2.1. Shared Battery Setting

The first case is rather simple. We consider a scenario where two or more households have access to PV generation and a shared battery. The battery can be charged or discharged at any given moment using either purchased or generated energy. Discharged energy can be sold to the retailer or used to satisfy their own energy needs. For simplicity, the power efficiency of this battery is high enough to make the energy loss from charging negligible. A visualization of such a community can be found in Figure 3.2. In a perfect scenario, each of the households would contribute to the shared energy pool by charging it with the excess of their produced energy and discharging it during periods of low energy generation or high energy prices. Moreover, since the costs of importing energy from the retailer greatly exceed the profits from selling the produced energy, the most efficient way of disposing of produced energy is storing it for later use.

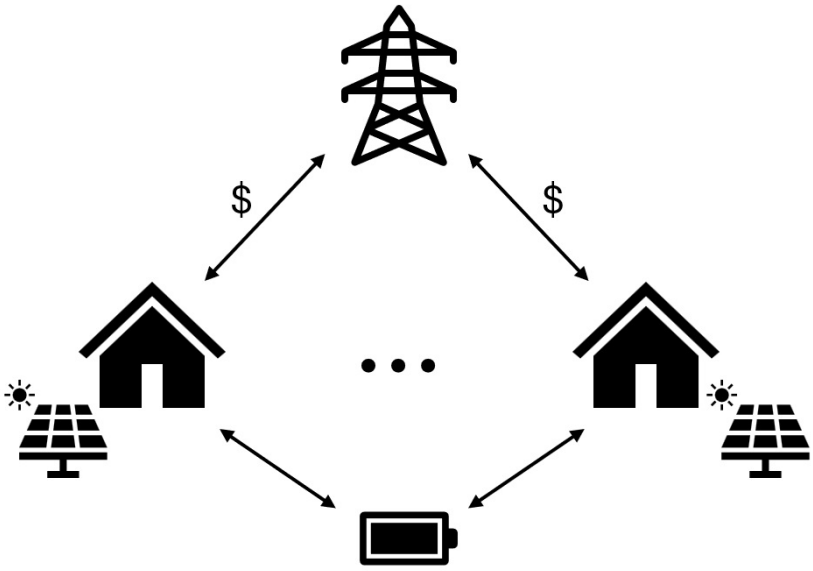


Figure 3.2: Shared battery setting. Note that this scenario allows for more than two households.

However, problems start to arise when some of the energy-sharing participants decide to exploit the system. Defiant households may decide to defect by refusing to contribute to the energy pool while continuing to draw energy from it, taking advantage of more good-willed households. If most of the households adopts this sort of behavior, the battery will lack the energy to benefit anyone. On the other hand, if only a handful of houses exploit the system, they make a lot of profit from selling their excess energy while consuming energy from shared resources. In the most extreme cases, some households may even decide to drain the shared battery entirely and sell all the discharged energy. This results in a social dilemma. In the next parts of the research, we will look closer at this scenario and try to re-create it in a reinforcement learning context.

3.2.2. Energy Trading with Unequal Storage and Generation Capabilities

A second scenario considers energy trading rather than shared batteries. Here, two households are present- Household A and Household B. Household A is capable of generating energy but has very inefficient energy storage - only a portion of the used energy can be retrieved from the charged batteries. By contrast, Household B has no energy generation capability but it is equipped with an extremely efficient energy storage that allows it to store energy with minimal energy loss. Moreover, a local energy market is present allowing households to exchange energy at any given time. For simplicity reasons, the energy is exchanged at a price equal to the mid-point between the current retail export and import prices ensuring that both parties can profit from such transactions. A visualization of this setup can be seen in figure 3.3. In this setup, it is generally more profitable for household A to store its energy locally rather than trade it with the retailer. However, for both groups to make the most profit, the cooperation is crucial. During peak generation periods, Household A can sell its excess energy

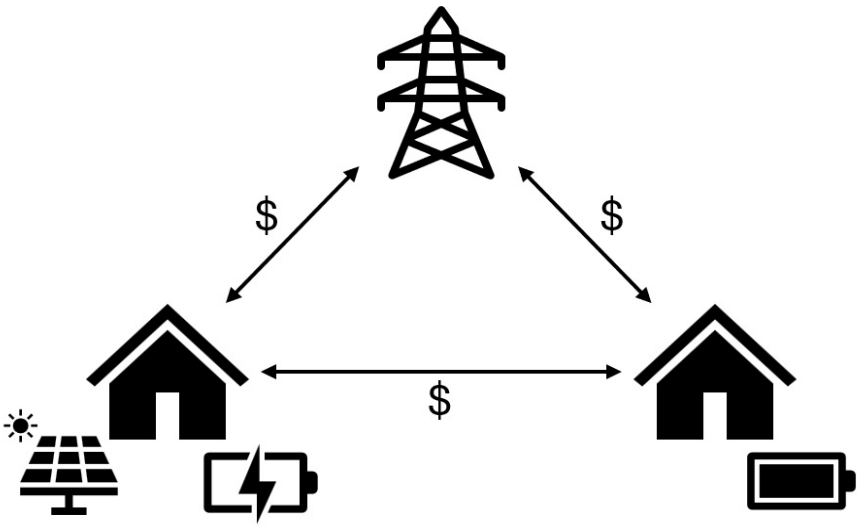


Figure 3.3: Scenario involving two households: one equipped with an efficient battery and the other with PV panels and an inefficient battery. Both households can trade with each other and the retailer.

to Household B which stores it efficiently. Household B is then expected to sell a portion of energy back to Household A so that Household A's net gain is higher than in the case of storing it inefficiently. Household B can still use a part of the stored energy for its own need, provided that Household A still benefits from the transaction.

Let us consider three possible scenarios:

1. Both households cooperate. Household B helps Household A store its generated energy and returns enough to make this transaction beneficial for both parties.
2. Both households refuse to cooperate. Household A stores all of its energy locally losing a huge portion of its generated energy due to inefficiencies. Household B gains nothing as it has no means to charge its local battery.
3. Household A decides to cooperate but Household B defects. In this scenario, Household A sends its energy to Household B hoping for a cooperative stance. Household B accepts the energy and stores it in its local battery. However, later on, Household B exploits this dynamic by refusing to trade back its stored energy and using it all to satisfy its energy needs making a huge profit but leaving Household A at a significant loss.

This scenario, though more convoluted than the first one, highlights the complex dynamics emerging in a seemingly simple environment yet still realistic environment. It is also a great example of why local energy communities provide a rich ground for exploring sequential social dilemmas.

4

Experiments

4.1. Experimental Procedure

To analyze how social dilemmas arise in the context of LEMs, we conducted a series of experiments focusing on minimizing the energy costs of the LEC. We chose social welfare, understood as the sum (or the average) of household rewards, as our guiding metric. The first experiment analyzed the training outcomes in an environment where no interactions between agents take place. The main goal was to assess how effectively agents can learn to store energy in a simplified setting. Following that, we introduced environments where social interactions take place through local trading and the use of a communal battery. By analyzing agent rewards and employing methods from existing literature, we investigate how social dilemmas arise under different training conditions. In the third experiment, we proposed a solution to help agents learn in such environments. Lastly, we investigated how these results translate to a complex and more realistic scenario involving several differing agents. A concise summary of experiments can be seen in Table 4.1.

Experiment	Description	Local Storage	Shared Storage	Mitigation	Trading
I	Basic agent capabilities	**Yes**	**No**	**No**	**No**
II	Identifying social dilemmas	**No**	**Yes**	**No**	**No**
III	Mitigation of social dilemmas	**No**	**Yes**	**Yes**	**No**
IV	Complex Setting	**Yes**	**Yes**	**Yes**	**Yes**

Table 4.1: Summary of experiments.

Some characteristics are shared across all the experiments we conducted. The used dataset was divided into two training and evaluation subsets with the training encompassing five and a half months out of the total six-month period. The evaluation set was carefully selected to include periods with varying power generation levels. Due to the length of a single simulation, it required around four training iterations employing several Ray workers to fully run it, contributing to abrupt jumps in the reward plots.

As outlined earlier, the PPO method was employed to train the algorithm. Hyperparameter values and deep learning model parameters were optimized via an extensive grid search and the detailed values can be found in Appendix A. Among these, the batch size, set to 50000, was one of the most critical factors. Smaller batch sizes failed to capture the variability of the environment causing the model to miss minima and coverage to sub-optimal solutions. Another significant enhancement that made the training feasible included the addition of reward shaping. During training, the agents were not rewarded

for selling the energy to the retailer. This prevented instant gratification and instead encouraged them to manage energy more strategically, promoting better utilization of storage systems.

4.1.1. Data Source

For energy load data and PV generation data, the environment relied on the Pecan Street dataset[96], one of the most widely used datasets for energy research and policy-making. More precisely, we utilized a subset of data collected in New York, which was enough to provide a vast array of energy behaviors. Examples of day-long energy generation and load can be found in Figure 4.1. While outside the scope of this research, the Pecan Street dataset also includes information about water, gas, and appliance-level energy usage, opening avenues for future research.

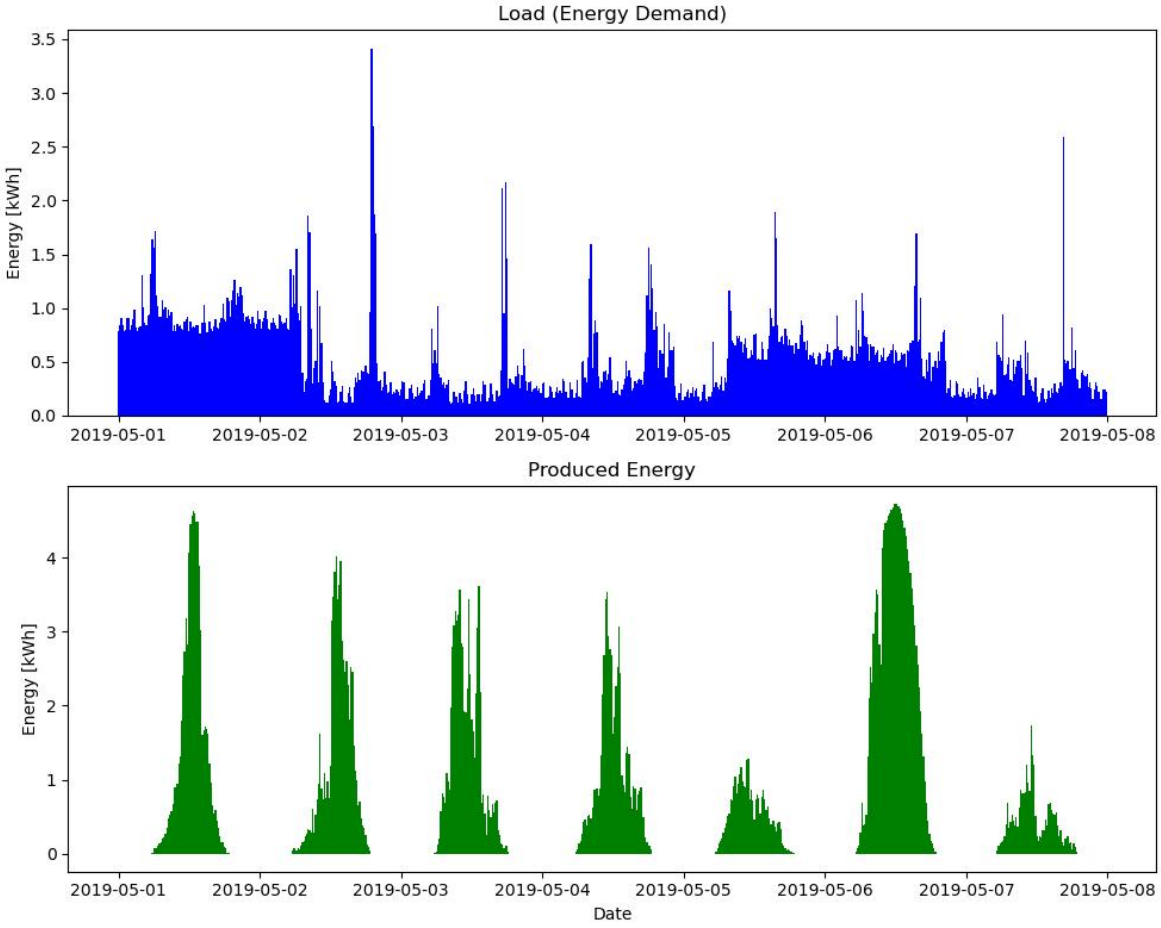


Figure 4.1: First week of the power generation and demand of the household type used in the first experiment. Each bar represents a 15-minute long time window.

4.1.2. Pricing

Each household is equipped with the ability to trade with an external retailer, albeit at rather unfavorable prices. Following common real-life energy contracts, Time-of-Use (TOU) pricing is in place, dividing the day into three periods with nighttime tariffs being much more advantageous for the customers. For export to the retailer, a fixed price is used throughout the day. Similarly to real-world energy contracts, this export price is notably lower than the retail import price. Figure 4.2 shows a plot containing daily export and import prices that are consistent across all households. Using TOU tariffs with realistic retail pricing ensures that our representation remains practical and resembles actual energy markets.

Before choosing Time-of-Use tariffs for pricing, wholesale energy pricing data from the New York Independent System Operator (NYISO)[97] was used. However, we opted for a different solution due to the

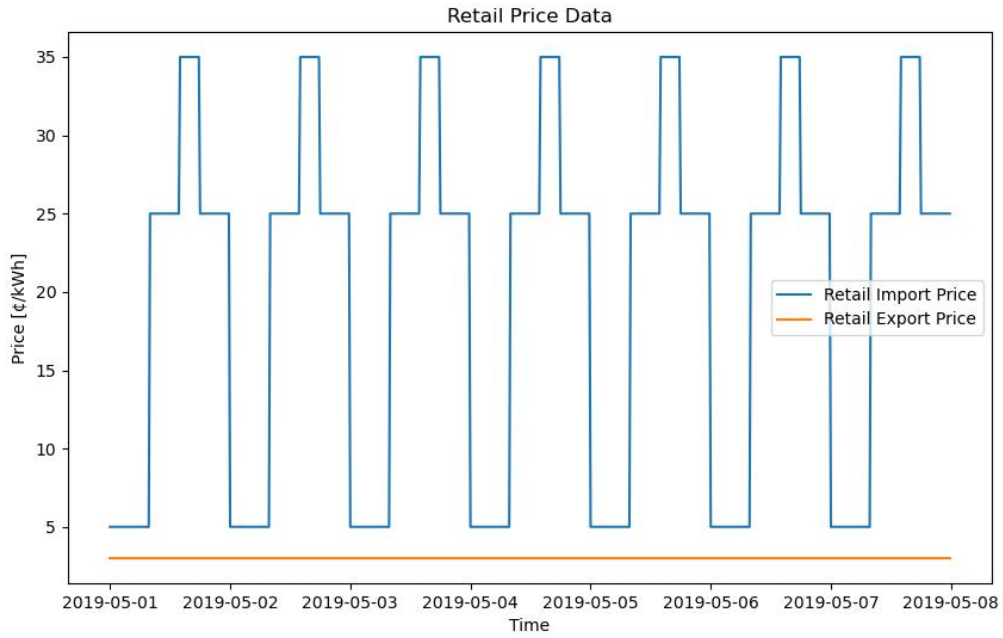


Figure 4.2: TOU pricing pattern for the first 7 days of the experiment, with energy prices peaking in the afternoon and reaching their lowest at night. The export price is equal to 3 cents per kWh at all times. While the plot only covers the first few days, the pattern remains consistent throughout the entire experiment.

high volatility of wholesale energy prices. Moreover, using wholesale instead of retail pricing created an unrealistic scenario but granted single households an opportunity at much lower prices than would be realistically possible driving the environment away from the real world.

4.2. Experiment I: Evaluating Basic Agent Capabilities

The first experiment focused on testing the environment and the basic learning abilities of the agents. To that end, we conducted a series of training and evaluation runs to test the capabilities of the agent with no inter-agent cooperation involved. This basic analysis is crucial for understanding agent dynamics when interactions are introduced in the other experiments.

We examined the environment's efficiency by running the training procedure on one type of household with moderate power generation capability, as depicted in Figure 4.1. The household was equipped with local battery storage that can be used to store produced and bought energy. We employed an ideal battery model, assuming that the energy can be stored and retrieved without any energy loss. This simplification helped with the reasoning about the optimality of such a system where storing generated energy is always beneficial compared to selling it to the retailer. To account for the inherent stochasticity of the training process, we repeated the experiments for ten households and averaged out the final rewards.

The reward plot can be seen in Figure 4.3. For both training and evaluation, a baseline reward was calculated based on a scenario where no energy is stored. In this scenario, solar energy production is used to directly satisfy the household's energy use, with energy excess sold and energy shortfall bought from the retailer. As you can see in the plots, the agents learn to operate in this sort of environment effectively. By looking at the energy plots from Figure 4.4 it is also evident that the agents successfully manage to store energy during peak power generation periods to then utilize it when prices remain high but the generation approaches zero.

To compare how agents behave in environments with varying resource availability, we conducted several training runs with different levels of power generation. These levels were obtained by scaling the original PV generation data by the PV generation coefficient η_{pv}. The comparison of evaluation rewards

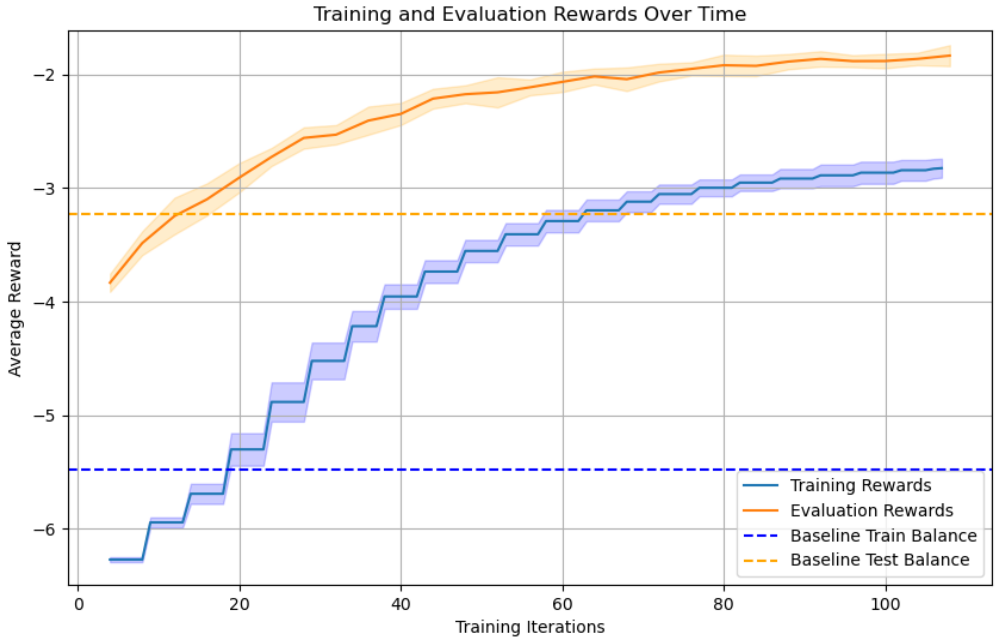


Figure 4.3: Reward plot from the first experiment. Each point represents the average reward from multiple simulation iterations across ten households. The shaded area indicates two standard deviations around the mean.

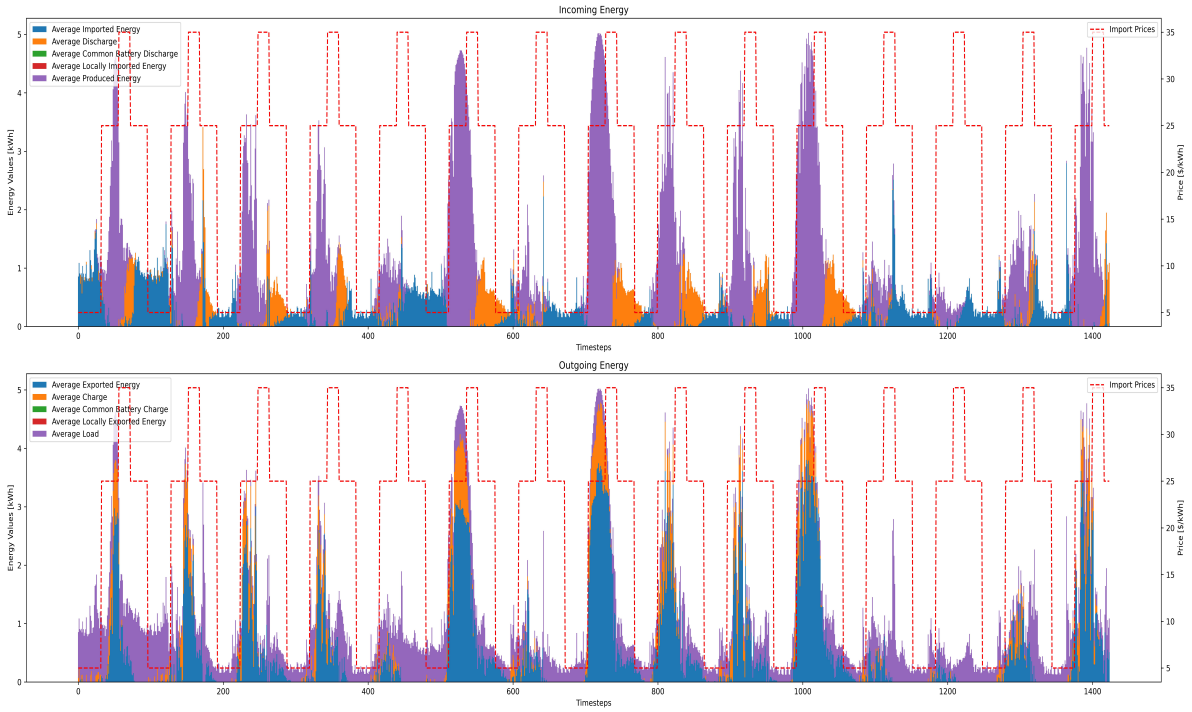


Figure 4.4: The distribution of incoming energy sources and outgoing energy destinations, averaged across ten identical households in a scenario with no interactions between agents. The energy generation coefficient is equal to 1.0. One timestep is equal to a 15-minute time window in the simulation. Energy stored and withdrawn locally is marked in orange.

for different resource levels is shown in Figure 4.5. The results clearly demonstrate that agents manage to exceed the baselines only in the two environments with the highest resource abundance.

The next step is to determine whether the agents fail to learn the intended strategic behavior or if the environment itself does not provide enough resources to outperform the baselines. Figure 4.6 confirms the former. The results indicate there exists a surplus of energy that is exported to the retailer during the peak production periods instead of being stored for later use which is by far a more beneficial behavior. This means that the agents struggle to adopt effective strategies in resource-scarce environments.

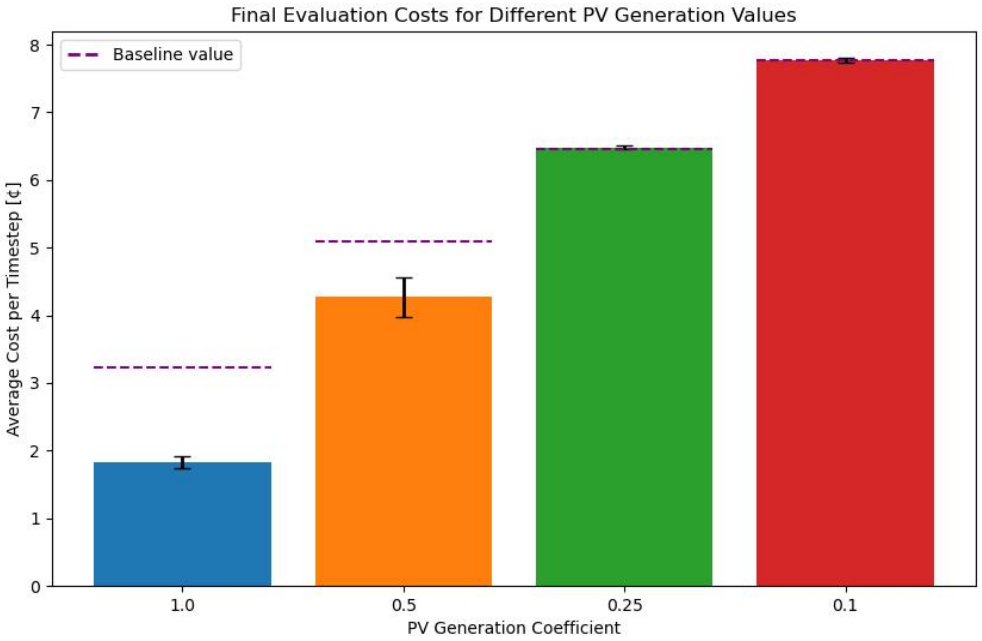


Figure 4.5: Evaluation rewards for groups of agents in environments with varying power generation coefficients η_{pv}.

4.3. Experiment II: Investigating Sequential Social Dilemmas

Having demonstrated that the agents can efficiently learn how to operate in an environment devoid of social interactions, we now turn our attention to investigating cases where the agents interact with each other. Specifically, we aim to investigate how social dilemmas arise in this sort of environment through the analysis of the learning outcomes and a matrix-based method described in the next subsection. This will also clarify why we previously focused attention on environments with varying resource availability.

4.3.1. SSD Detection Procedure

To identify and analyze the presence of SSDs in our environment, we follow a procedure inspired by the work of Leibo et al [9]. As established earlier, resource availability has a very strong influence on the efficiency of learning within the LEC environment. Here, we assume that this trend is equally present when inter-agent interaction is involved which holds in many other MARL environments. Under this condition, we can match agents characterized by varying cooperativeness due to them being trained in environments with different resource scarcity. This allows us to observe the rewards of agents in environments where all agents cooperate, some agents exploit others or no cooperation occurs. These rewards are exactly the four payoff values, as discussed in Subsection 2.2.

Following this direction, we begin by collecting two sets of policies Π^A and Π^S that were trained in environments with abundance and scarcity of resources, respectively. Next, we run the simulation with agents characterized by varying tendencies for cooperation. From policy sets, we sample two pairs of policies (π_1^A, π_1^S) and (π_2^A, π_2^S) which are then matched against each other within the investigated environment. The matches are played in four distinct configurations, as illustrated in Figure 4.7. This way, we derive the values for R, S, T, P (see Subsection 2.2) for the given configuration of policies, which, in turn, allows us to calculate the fear and grid factors influencing the agents' actions. A visualization

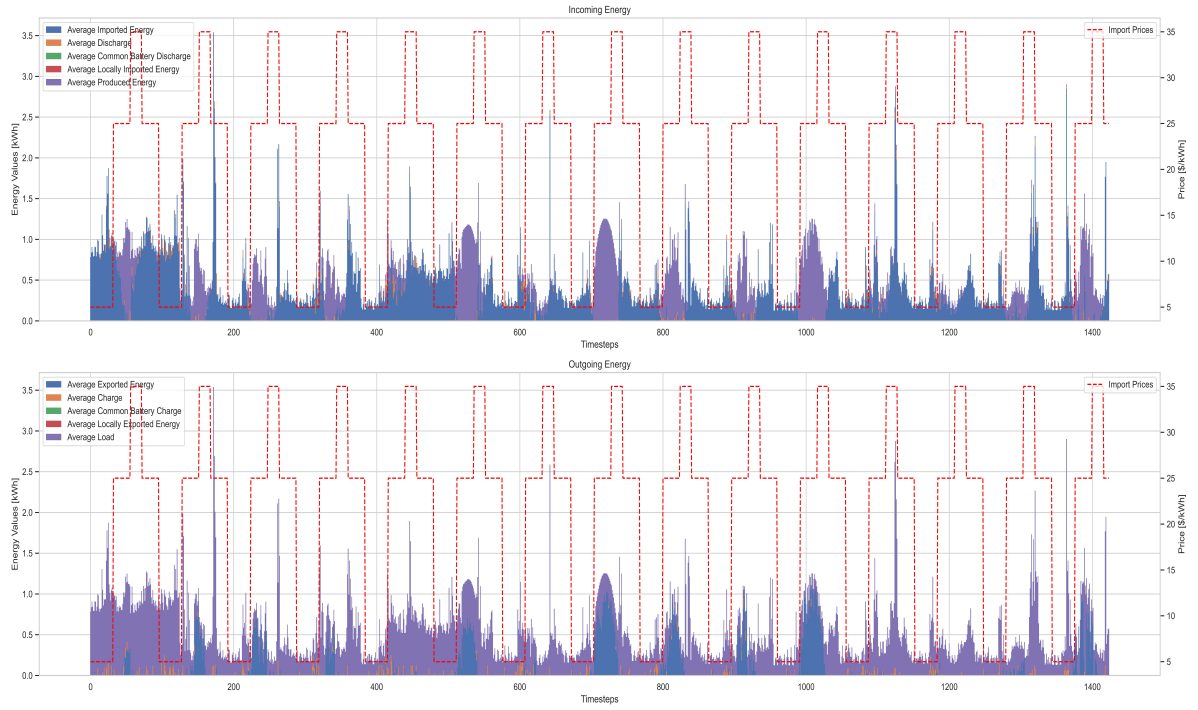


Figure 4.6: The distribution of incoming energy sources and outgoing energy destinations, averaged across ten identical households in a scenario with no interactions between agents. The energy generation coefficient is equal to 0.25.

of one iteration of the described procedure can be seen in Figure 4.8.

Several evaluation rounds take place for each matching until the values converge. To collect enough data samples, it is then repeated for other policy samples. This way, we can develop a solid understanding of the kind of social dilemmas present in the environment. Aside from being an indicator of the presence of SSDs, this analysis helps us evaluate the underlying motivations for the agent's defiant behavior- whether it is guided by fear of exploitation, greed for exploitation, a combination of both, or neither. Moreover, it situates each configuration of policies into one of four quadrants corresponding to the games described earlier, further clarifying the dynamics at play.

	Cooperate	**Defect**
Cooperate	(π_1^A, π_2^A)	(π_1^A, π_2^S)
Defect	(π_1^S, π_2^A)	(π_1^S, π_2^S)

Figure 4.7: Policy pairings for matching. The mutual abundance scenario is highlighted in green, the mutual scarcity scenario in red, and the mixed scenario—where households were trained in both scarce and abundant environments—is highlighted in orange.

4.3.2. Case with Communal Storage

In this part of the experiment, we aim to recreate the scenario described in the Subsection 3.2.1. The environment is set up similarly to the one in the first experiment. Except, this time, the agents do not have access to personal batteries and can operate on a communal battery instead. Again, we assume an ideal charging and discharging model with $\eta_{\text{ch}} = 1.0$. There are 3 households in every environment. To provide a fair ground for later matches, we ensure every household is equal in terms of their generation capabilities. Thus, the power generation curves are identical for each household.

We completed a full training and evaluation procedure ten times for each of the two values of the PV generation coefficient η_{pv} - 0.5 and 3.0- representing scarce and abundant environments which we refer to as 'scarce' and 'abundant' policies respectively. This gave us two sets, with 30 policies in each set,

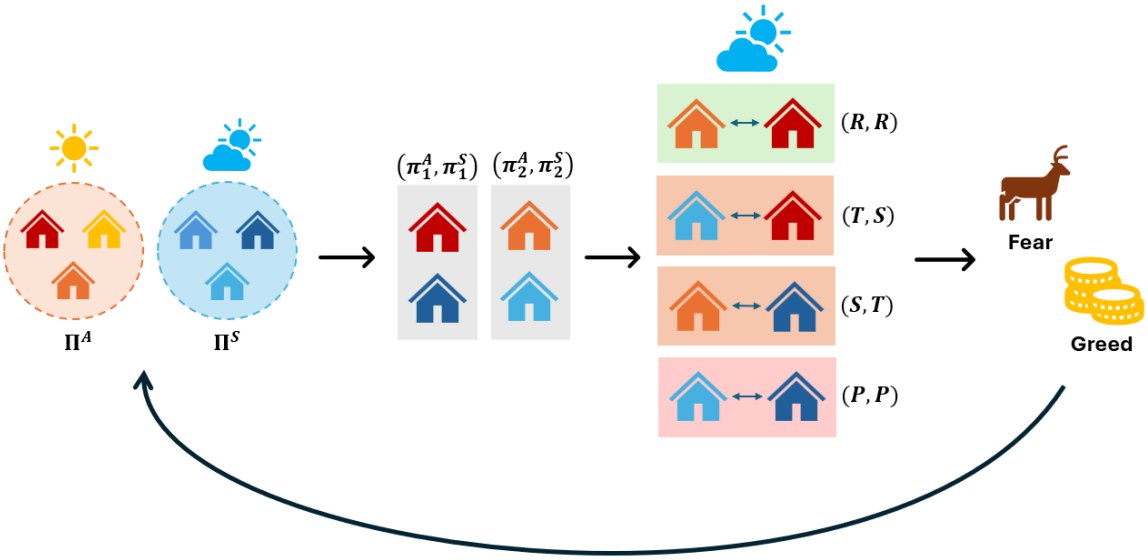


Figure 4.8: Visualization of the Sequential Dilemmas detection procedure in the LECs environment. First, two sets of policies Π^A and Π^S are trained on resource-rich (warm hues) and resource-scarce (cold hues) environments. Then, two pairs of policies are sampled from both of these sets and matched against each other. Lastly, the fear and greed factors are calculated based on the results of these matches in four different configurations. This iterative process is repeated several times and results from each iteration are plotted. Examples of such final plots can be seen in Figure 4.11.

trained in environments with different levels of resource availability. Following that, we matched them as described in Subsection 4.3.1.

Before we begin analyzing occurring SSDs, we need to confirm our initial assumptions by assessing the differences in agents' cooperativeness across environments with different resource scarcity. We start by comparing the evaluation rewards across the three used matching configurations - *abundant* where both policies were trained under abundant conditions, *scarce* where both policies were trained under scarce conditions, and *mixed* where policies were derived from both conditions. All matches took place on an identical scarce environment with $\eta_{\text{pv}} = 0.5$. The scatterplot in Figure 4.9 shows the difference in evaluation rewards between the three configurations. The results make it evident that the policies trained under abundant conditions obtain higher final rewards compared to those in both the mixed[1] and scarce environments[1]. Furthermore, the mixed environment scores higher than the scarce one [1]. This finding is particularly interesting as it demonstrates that abundant policies outperform scarce ones, even in environments with the same resource scarcity as the training conditions of the scarce policies.

To better understand the root cause of reward differences, we now examine the mixed environment. This setting is especially illustrative as it highlights behavioral distinctions between the two policy sets and reveals variations in the levels of selfishness among agents trained under different conditions. Figure 4.10 depicts the rewards obtained by agents in this configuration. The results reveal that abundant policies generate lower overall profit (measured as the net difference between export profits and import costs) compared to their scarce counterparts [1]. This is most likely due to the fact that scarce policies exhibit more aggressive behavior. Specifically, when interacting with the shared battery, they discharge more and charge less than agents with abundant policies [1].

These discoveries suggest that agents following abundant policies are significantly more cooperative than those trained under scarcity conditions, resulting in a greater cumulative reward in environments where both agents adopt this cooperative strategy. Therefore, our first key takeaway from the experiments is that rescaling the training data can help mitigate social dilemmas in the LEC setting. By improving resource scarcity during training, we bolstered cooperative behaviors in the agents leading to an increased social welfare of the whole community.

Reassured that our initial assumption holds, we can now analyze the occurrence of SSDs based on

[1] p-value < 0.001, Test Used: Mann–Whitney U

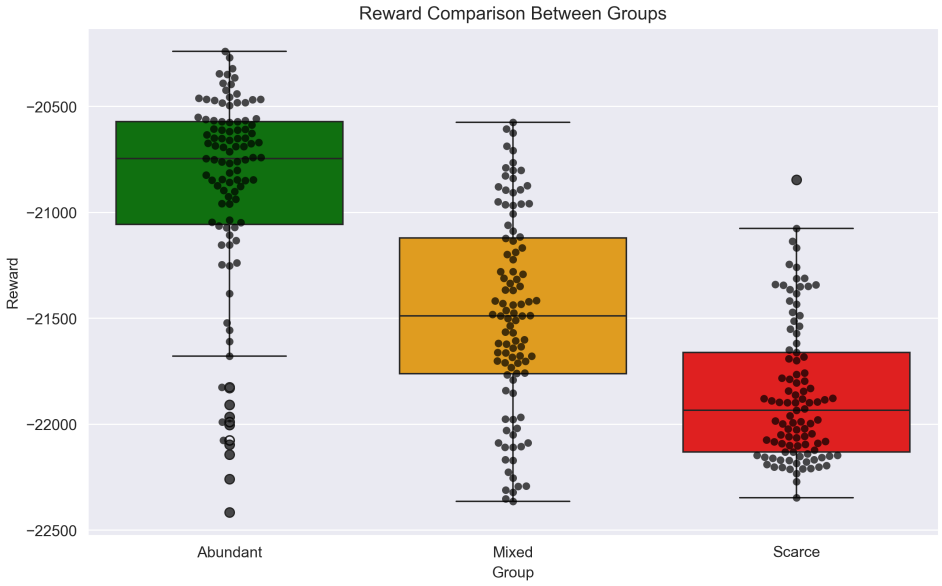


Figure 4.9: Comparison of averaged evaluation across three matching configurations. Rewards for each configuration were calculated from 100 averaged matches, with each match consisting of several evaluation runs between two policy pairs. In the matching environment, η_{pv} was set to 0.5.

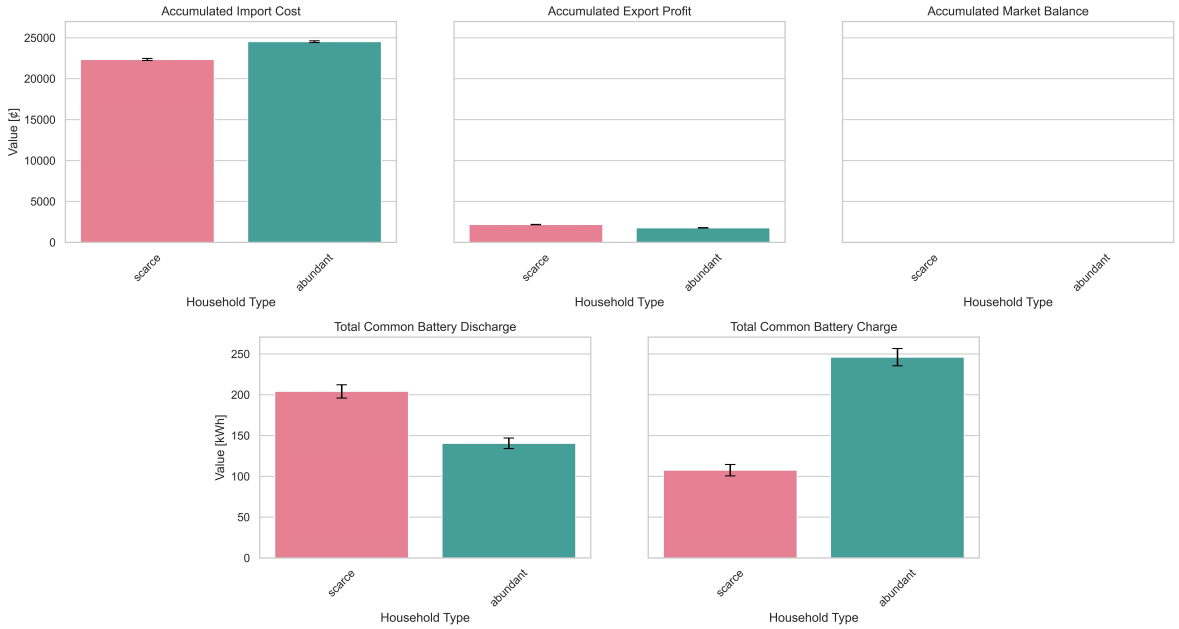


Figure 4.10: Comparison of key energy metrics across two distinct policies, including import costs, export profits, market balance, and shared battery usage (charge/discharge). The abundant policy was trained in an environment with six times the energy generation of the scarce policy. Metrics are averaged over 200 policy pairs sampled from 30 distinct policies. Black error bars mark the 95% CI. The accumulated market balance is not shown, as trading is disabled in this environment.

the outcome of the matches. To develop an even deeper understanding of how resource scarcity influences the scale of SSDs, we performed an additional round of matches on the environment with $\eta_{\text{pv}} = 1.0$. Figures 4.11 illustrate the final fear-greed ratios in both of the employed environments. Each dot on the plots represents the outcomes of matches played between agents following policies from two distinct policy sets. The results lead to two noteworthy conclusions. First, in both cases, the match outcomes span all three SSD classes. That means that the SSDs indeed occur in the shared battery environment and are driven by both fear and greed. This finding supports the hypothesis outlined in the

subsection 3.2.1. Second, the availability of resources significantly influences the prevalence of SSDs. In other words, agents are more prone to adopt defective stances in environments that are scarce in resources. This is intuitive as the scarcity of resources amplifies the potential benefits of exploitative actions, making them more attractive to agents.

4.4. Experiment III: Mitigating Sequential Social Dilemmas

As our findings have shown, the social dilemmas of a sequential nature are indeed present in our environment leading to suboptimal community outcomes if some agents decide to exploit their more cooperative neighbors. We have also presented how transforming the training data can improve agents' willingness to cooperate. In the following section, we explore another modification that can be applied to the learning procedure to reduce such exploitative behavior and improve the social welfare of the whole community. Specifically, we investigate two simple reward function modification in a setting that involves the use of common batteries.

Our previous observations revealed that houses that are more prone to positively contribute to the common battery, on average score much higher than those that frequently withdraw energy. Therefore, the modification we suggest incentivizes this sort of behavior by rewarding agents for charging the common battery and penalizing them for discharging it.

4.4.1. A Fixed Common Battery Reward Component

We start with a simple approach of adding a fixed incentive for charging the battery and a penalty for withdrawing the energy from it. It is done by introducing a new component to the reward function that is proportional to the amount of charged or discharged energy and is scaled by a fixed common storage reward factor β_s. The modified reward function takes the form:

$$r_t = P_{\text{im}}^{\text{retail}} \cdot E_{\text{im},t}^{r} + P^{\text{local}} \cdot \left(E_{\text{im},t}^{l} - E_{\text{ex},t}^{l}\right) + \beta_s * (E_{\text{ch}}^{s} - E_{\text{dis}}^{s}) \tag{4.1}$$

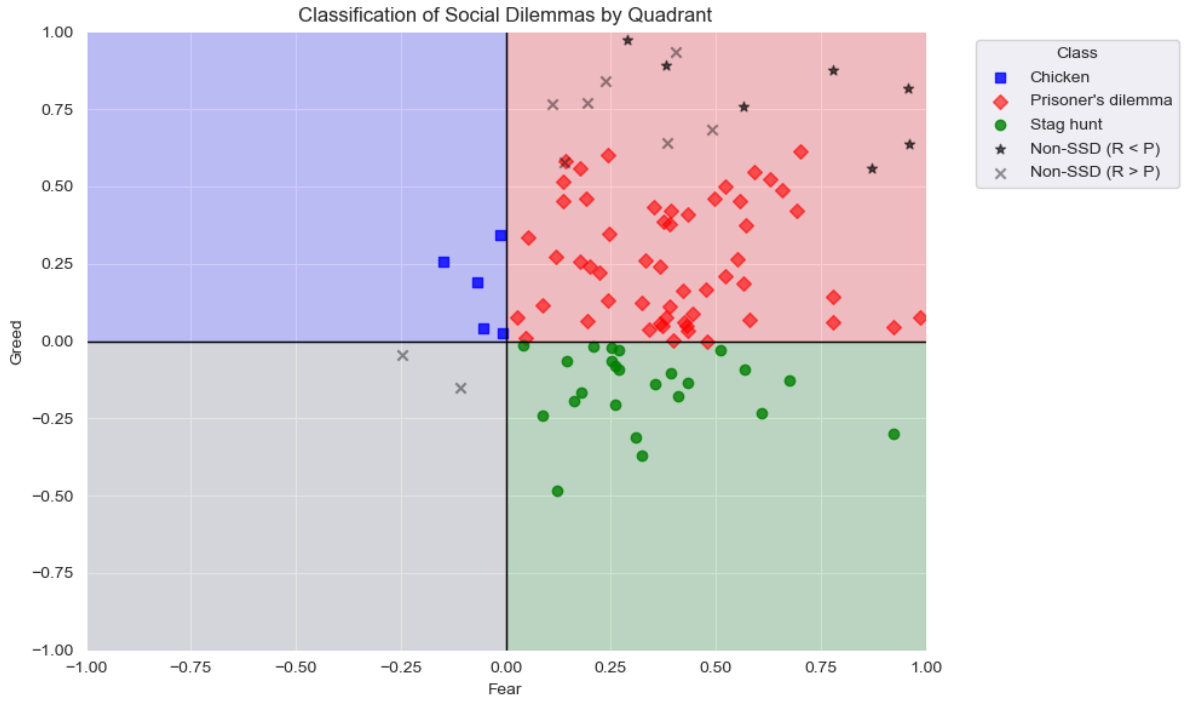
To evaluate the efficiency of this reward modification, we trained the agents in a resource-scarce environment ($\eta_{\text{pv}} = 0.5$). As demonstrated by previous experiments, this environment is especially well-suited to these experiments since agents found it difficult to learn cooperative behavior in this setting. The employed setup is identical to the one in the Experiment II. The training and evaluation procedure were performed using five different values of the β_s coefficient: 0, 1, 2.5, 5, 10, and 25. For each value, the procedure was repeated 10 times.

The plot with the final rewards is presented in Figure 4.12. The results confirm that incorporating the modified reward function improves the performance of the agents in our environment. Among the selected values, $\beta_s = 10$ yielded the highest average reward, improving the average reward by 23.67% compared to the original reward function. Values of 5, 2.5, and 1 also improved the results, albeit to a lesser extent. Interestingly, increasing β_s to 25 led to diminishing returns, performing worse than the baseline. A comprehensive comparison of the rewards along with their statistical significance can be seen in the table 4.2.

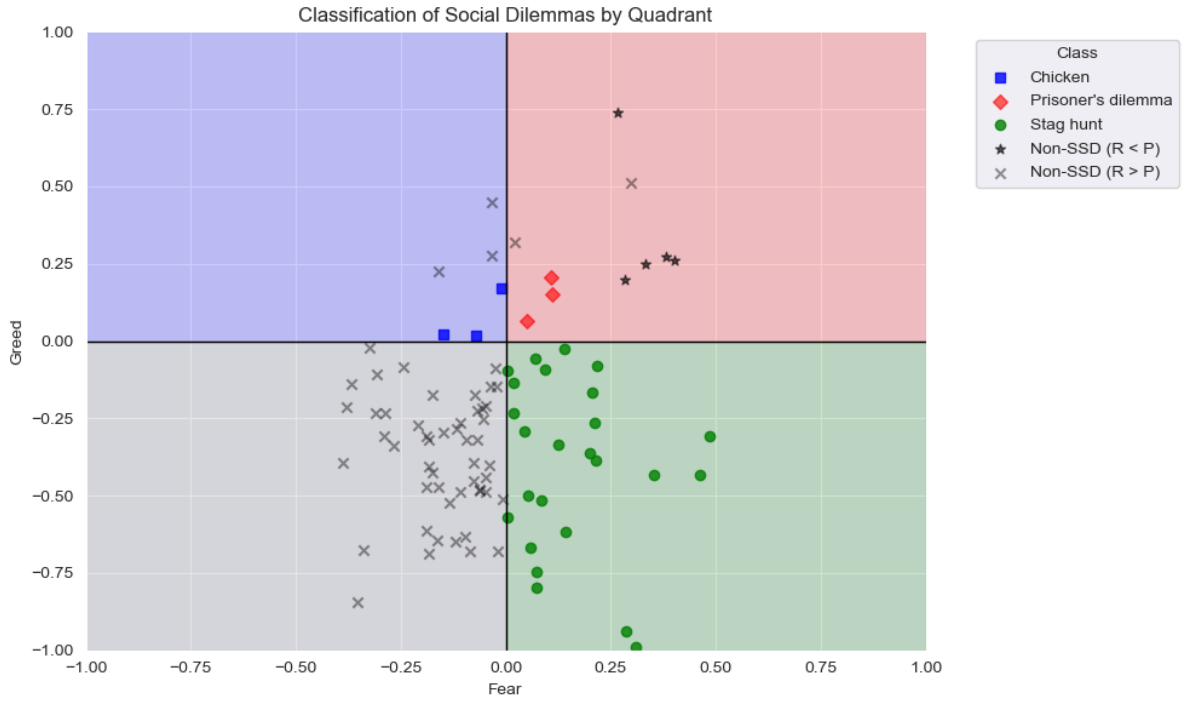
Coefficient	Group Mean	Difference	% Change	p-value
1.0	-14.9138	0.4602	2.99%	$p < 0.001$
2.5	-14.4287	0.9453	6.15%	$p < 0.001$
5.0	-13.1808	2.1932	14.27%	$p < 0.001$
10.0	-11.7357	3.6383	23.67%	$p < 0.001$
25.0	-15.6743	-0.3003	-1.95%	$p < 0.001$

Table 4.2: Pairwise analysis of rewards for different β_s coefficients relative to the reward with $\beta_s = 0$. The baseline mean was constant at -15.3740 across comparisons. The p-values were derived using the Mann–Whitney U test.

To investigate whether the increase in reward can be attributed to the greater cooperative behavior of the agents, we analyze energy flow patterns under different β_s values. Figure 4.13 depicts the energy flow when no reward modifications is applied ($\beta_s = 0$) while Figure 4.14 shows the cases for $\beta_s = 10$ and $\beta_s = 25$. Based on these plots, we can clearly see that the quantity of energy contributed by the



(a) The matches took place in an environment with $\eta_{\text{pv}} = 0.5$.



(b) The matches took place in an environment with $\eta_{\text{pv}} = 1.0$. That is, the households were able to generate twice as much energy compared to the case 4.11a.

Figure 4.11: Quadrant-based classification of social dilemmas occurring in the LECs environment, visualizing fear and greed metrics across three different game types. Each dot represents two pairs of policies matched against each other in four different configurations. The values have been normalized to the [-1; 1] range.

agents is much higher for $\beta_s = 10$ compared to the case where $\beta_s = 0$ throughout the whole evaluation period. This indeed demonstrates a clear shift towards more cooperative behavior encouraging them

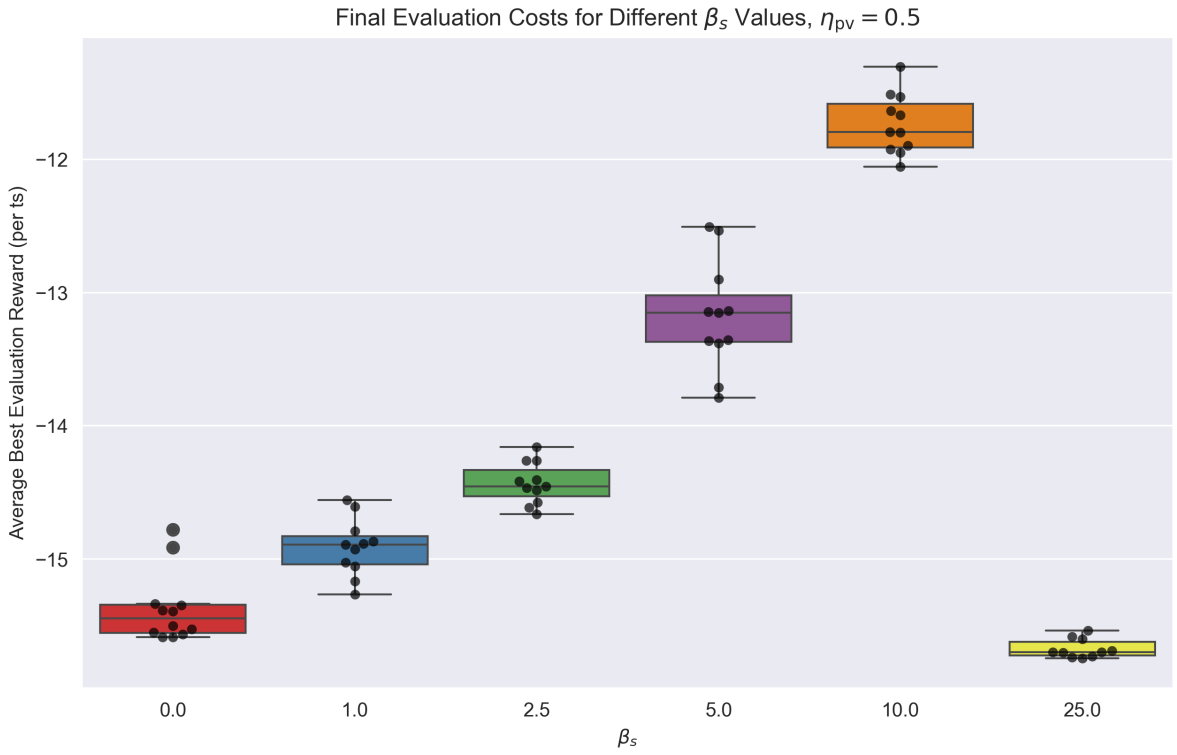


Figure 4.12: Boxplots with evaluation rewards for groups of agents in environments with different values of β_s.

to charge the battery and discouraging the from discharging it too excessively. However, looking at the plot for $\beta_s = 25$, we can observe that too large values of β_s lead to agents losing track of their energy-saving goal in favor of maximizing the amount of energy that is immediately channeled into the common batteries. In turn, this leads to highly suboptimal results. Therefore, these findings highlight the importance of tuning β_s appropriately to pick a value that encourages cooperation while maintaining agents' focus on their primary goal.

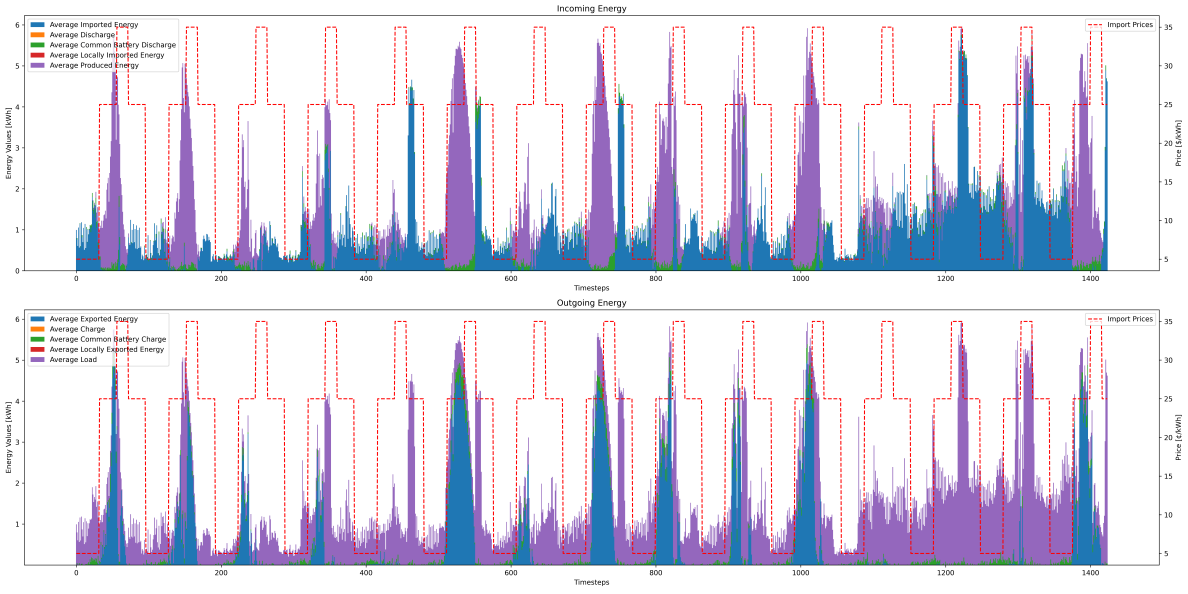
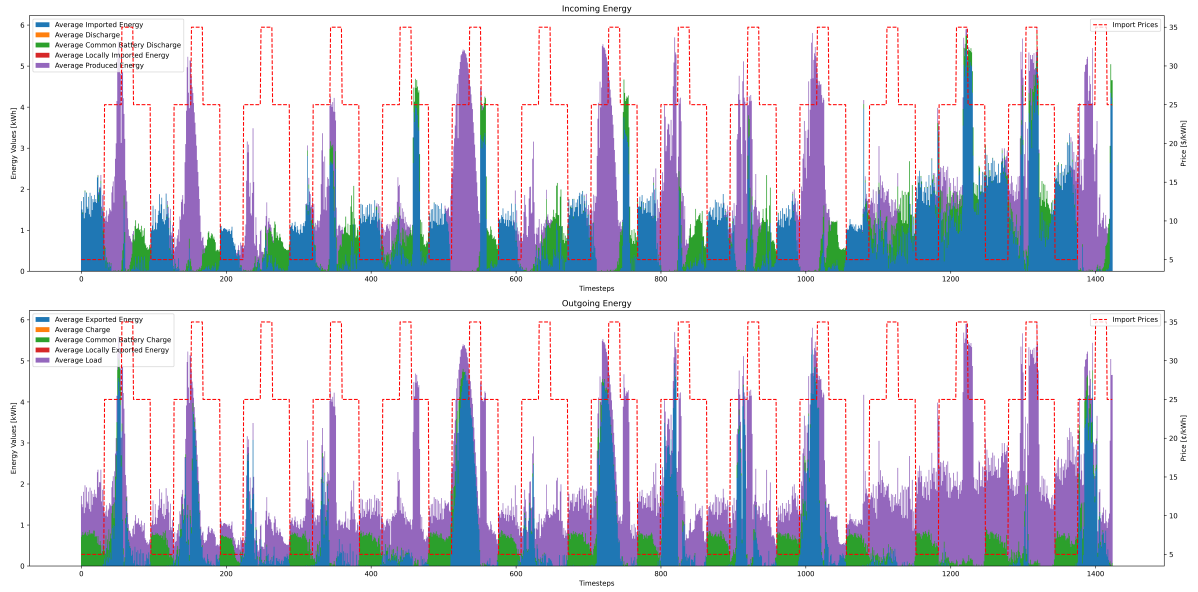
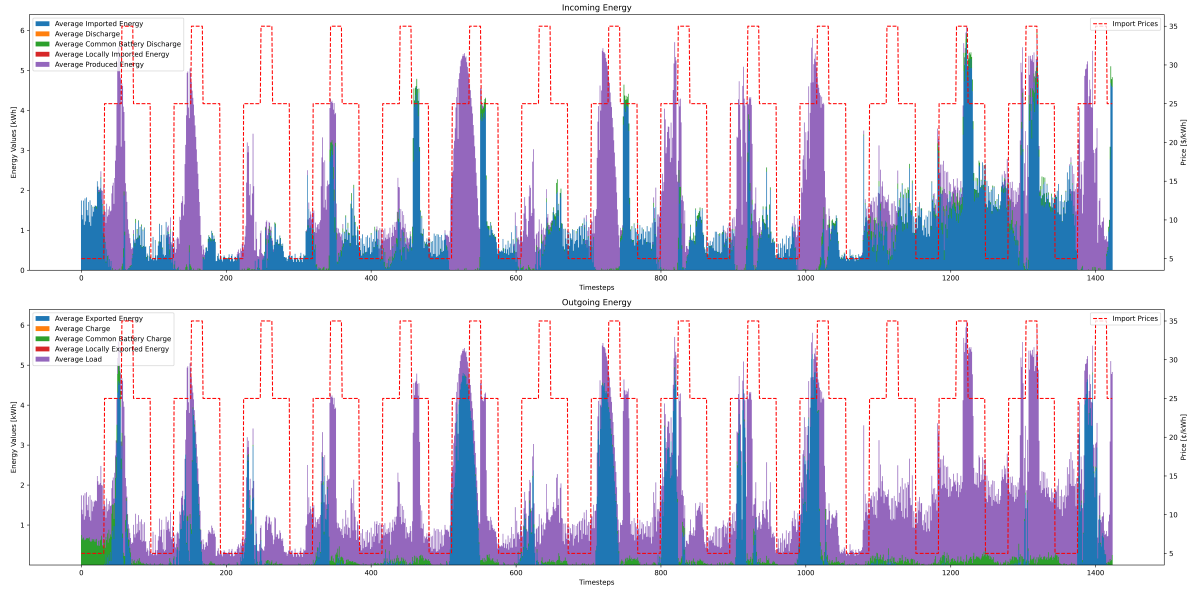


Figure 4.13: The distribution of incoming energy sources and outgoing energy destinations in a shared battery scenario for $\eta_{pv} = 0.5$. No modifications were applied to the reward function.



(a) $\beta_s = 10$



(b) $\beta_s = 25$

Figure 4.14: The distribution of incoming energy sources and outgoing energy destinations in a shared battery scenario for $\eta_{\text{pv}} = 0.5$. A fixed common battery reward alteration is applied to the reward function.

4.4.2. A Price-based Common Battery Reward Component

As an alternative to the modification presented in the previous section, we also evaluate the price-based incentive with a price-based factor λ_s. Unlike the fixed variant, the new price-based component of the reward function is additionally proportional to the current retail energy price. This way, we can discourage agents from withdrawing the energy when the energy is most valuable and incentivize them to contribute instead. This idea was guided by the assumption that the impact agents have on their environment when interacting with the battery is proportional to the energy's value at a given point of time. Thus, including this as a factor in the agent's reward could bring even more increase to the agents' final reward. The modified reward function takes the following form:

$$r_t = P^{\text{retail}}_{\text{im}} \cdot E^{r}_{\text{im},t} + P^{\text{local}} \cdot \left(E^{l}_{\text{im},t} - E^{l}_{\text{ex},t}\right) + \lambda_s * P^{\text{retail}}_{\text{im}} * \left(E^{s}_{\text{ch}} - E^{s}_{\text{dis}}\right) \tag{4.2}$$

The final outcomes of the experiments are illustrated in Figure 4.15 with more details present in Table 4.3. Introducing the price-based reward component results in an improved final reward with the $\lambda_s = 0.5$ resulting in the most significant reward improvement of 9.54% over the original reward function. However, the results do not suggest that the price-based component performs better than the fixed one.

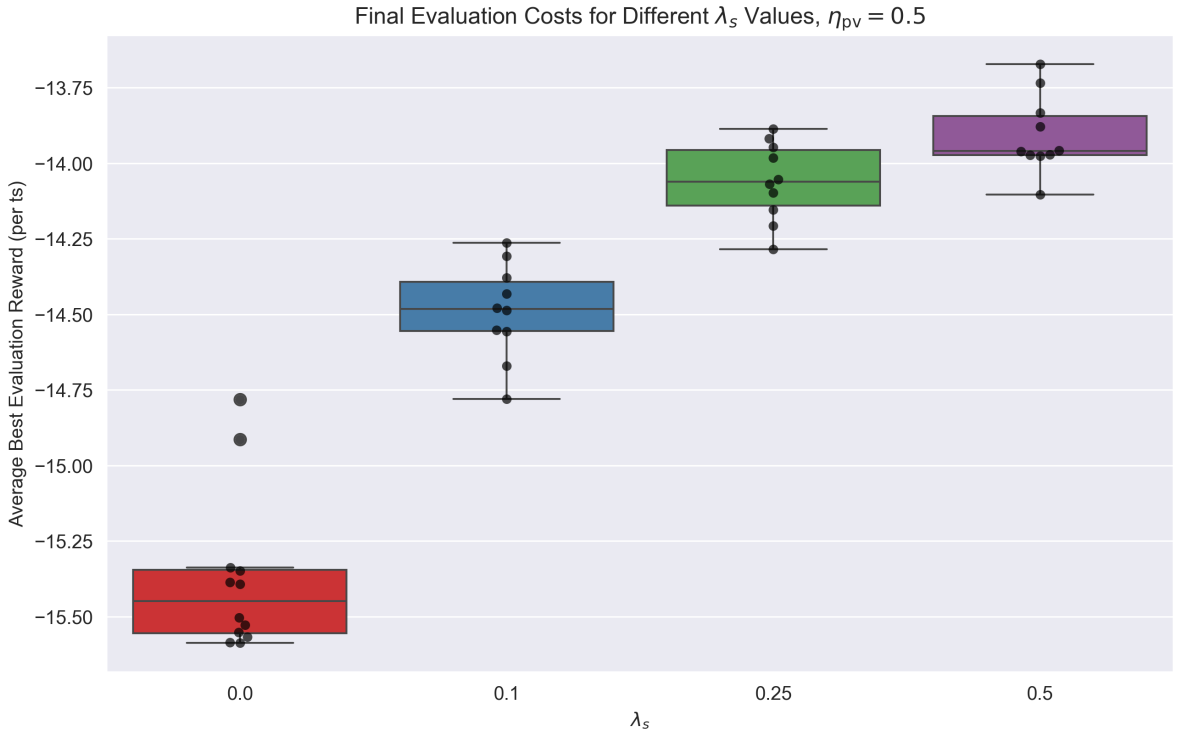


Figure 4.15: Boxplots with evaluation rewards for groups of agents in environments with different values of λ_s.

Coefficient	Group Mean	Difference	% Change	p-value
0.1	-14.4910	0.8830	5.74%	$p < 0.001$
0.25	-14.0606	1.3134	8.54%	$p < 0.001$
0.5	-13.9066	1.4674	9.54%	$p < 0.001$

Table 4.3: Statistical comparison of the final evaluation rewards for different λ_s price-based coefficients. The baseline mean was constant at -15.3740 across comparisons. The p-values were derived using the Mann–Whitney U test.

4.5. Experiment IV: Realistic Scenario

In this final experiment, we aim to evaluate how our findings translate to a more complex and realistic LEC environment. To achieve this, we constructed an environment consisting of 10 households each with varying energy needs and unequal access to energy storage and generation capabilities. Specifically, half of the houses are equipped with their own local energy storage, while the other half relies solely on shared storage. The houses were chosen so that the total energy generation of both groups was approximately equal. We abandon the lossless model of the batteries in favor of a charging efficiency model- a part of the energy is lost when charging the energy. The energy efficiency of shared and local batteries is 0.95 and 0.85 respectively. This means that, even though some of houses are equipped with their own battery storage, they can still benefit from the more efficient shared batteries.

Moreover, trading mechanisms are implemented allowing agents to exchange energy with one another as outlined in Subsection 3.1.1.

The objectives of this experiment are twofold. First, we examine how resource availability affects the training outcome. Our goal is to verify if, similarly to the cases considered so far, agents find it generally more difficult to learn in resource-scarce environments. For this purpose, we assess the agents' performance on environments with varying values of η_{pv}: abundant with $\eta_{\text{pv}} = 2.0$, normal with $\eta_{\text{pv}} = 1.0$ and scarce with $\eta_{\text{pv}} = 0.5$. Second, we assess whether the fixed reward function modifications (Equation 4.1) from the Experiment III yield similar improvements in this more complex scenario. We consider a fixed common battery reward with $\beta_s = 10$ as prior experiments identified this as the most effective value.

Figure 4.16 depicts the change in agents' performance relative to their respective baseline values, with a detailed breakdown of the outcomes in Table 4.4. Among the three resource availability scenarios, the agents only managed to exceed the baselines under abundant and normal resource availability conditions, while under scarce conditions, their performance fell below the baseline. These observations confirm that similarly to the previous cases, agents find it generally more difficult to learn on scenarios characterized by low resource availability. Given the low performance in the scarce scenario, the mitigation was applied during the training performed in this setting. With mitigation in place, the agents' training results improved substantially, outperforming the baseline by 10%. This suggests that our mitigation solution can be effectively extended to more complex cases. Moreover, these results show that, by employing the suggested training techniques, agents can indeed learn to effectively learn to operate in our custom environment. Most importantly, they are also capable of learning coordinated behaviors that require coordination between 10 different households.

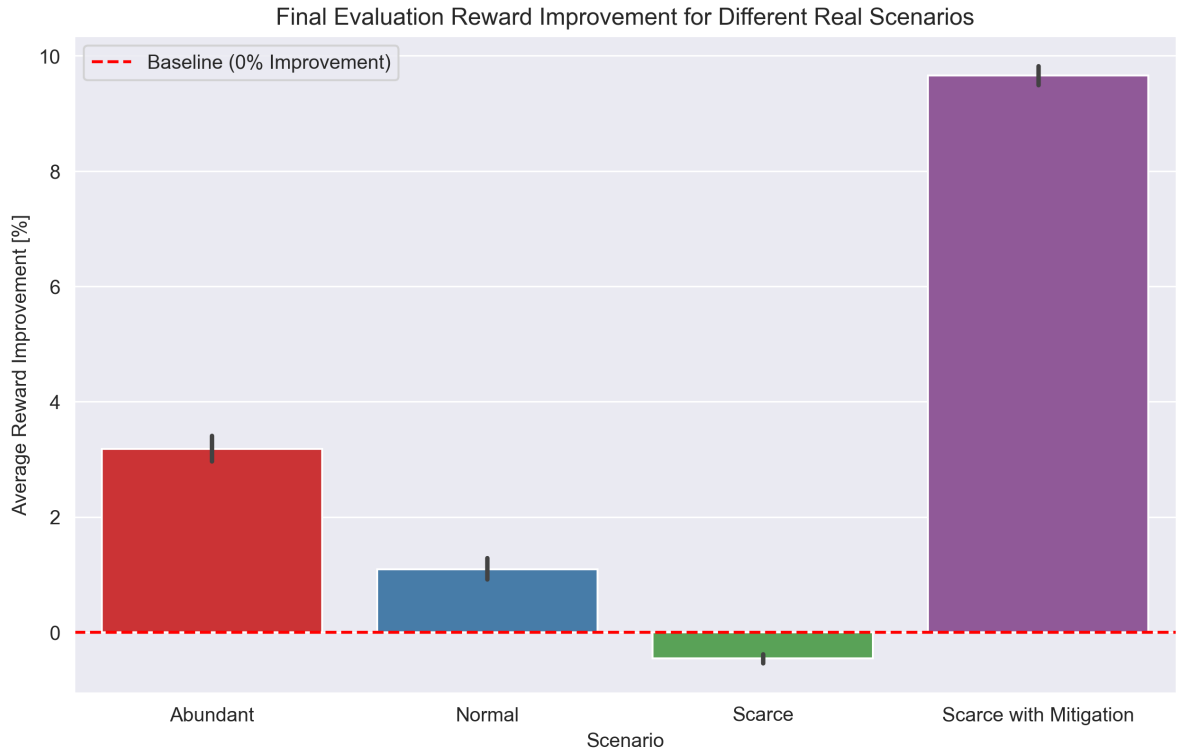


Figure 4.16: Final evaluation rewards for complex scenarios with different resource scarcity levels and a mitigation technique applied. The mitigation involves applying a fixed common battery component to the reward with $\beta_s = 10$. 95% CI is marked with dark bars.

To gain deeper insights into the impact our mitigation technique has on the trained policies, we now present a qualitative analysis of the energy plots from the two experiments conducted in the scarce environments (see Figure 4.17). One interesting conclusion is that, contrary to our initial expectations,

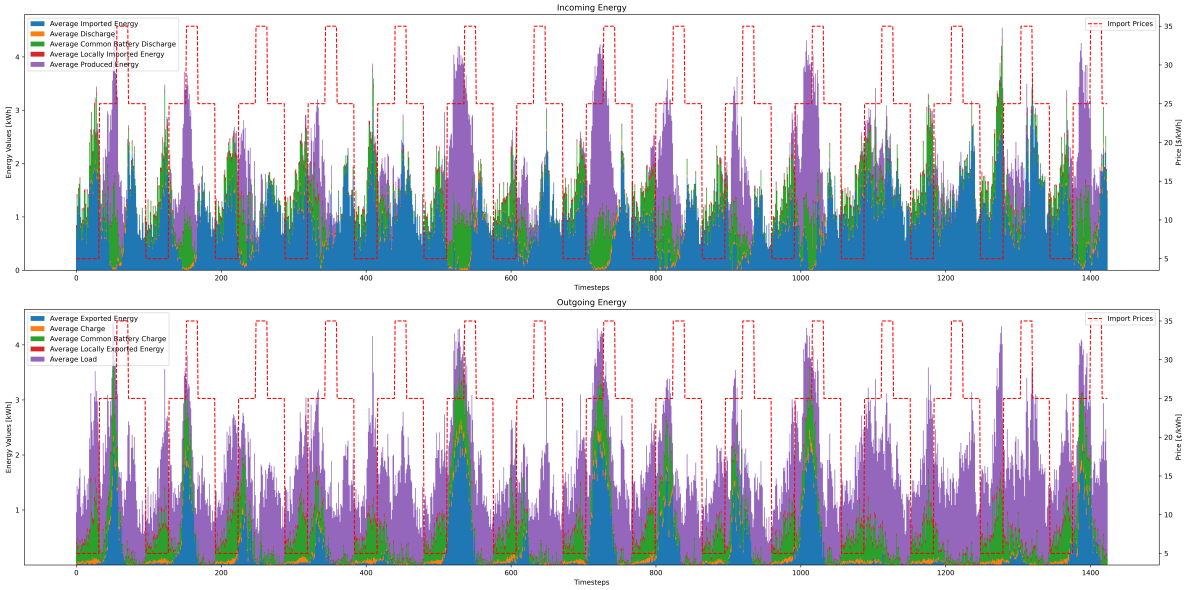
Scenario	Baseline	Group Mean	Difference	% Change	p-value
Abundant	-8.539	-8.268	0.271	3.18%	$p < 0.001$
Normal	-13.297	-13.151	0.146	1.10%	$p < 0.001$
Scarce	-16.591	-16.667	-0.076	-0.46%	$p < 0.001$
Scarce with Mitigation	-16.591	-14.989	1.602	9.66%	$p < 0.001$

Table 4.4: Statistical analysis by a scenario with improvement over a baseline specific to the given resource availability. The tests were performed using a one-sample t-test.

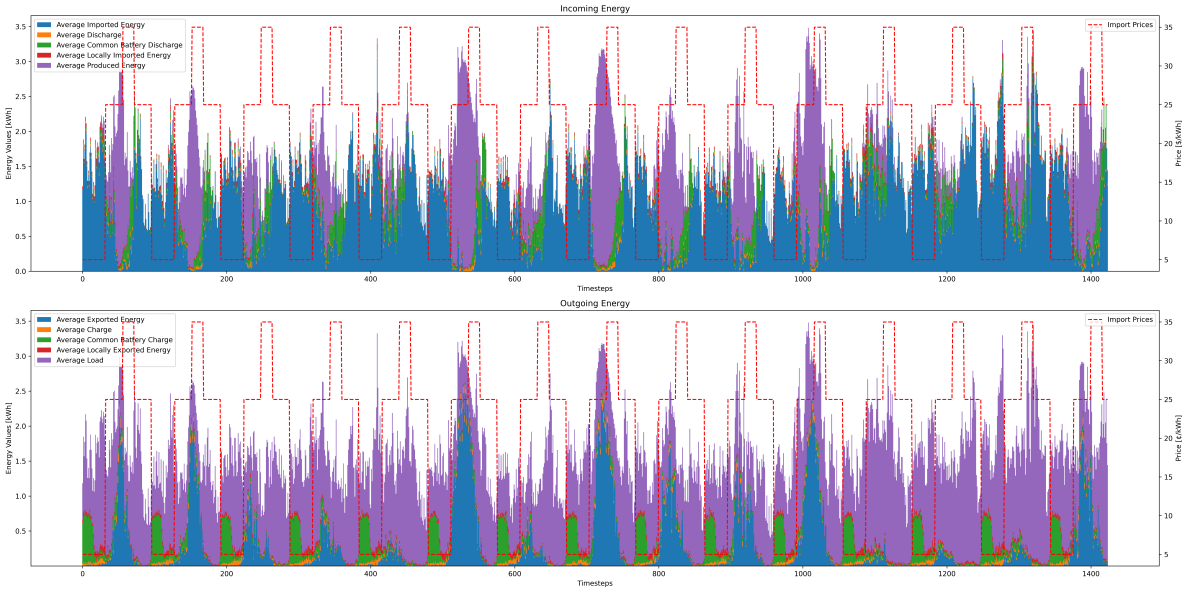
increasing the value of β_s does not result in increased use of the common battery in this setting. Instead, agents learn how to operate the battery in a highly organized manner - they charge it when the energy costs are the lowest and discharge it gradually over time when prices are higher and the generation remains low.

Another observation is that agents utilize their local batteries to a much lesser extent compared to the common battery which is logical given the difference in efficiency and the fact that only part of them are equipped with local storage. Finally, while some trading actions are present, only small quantities of energy are exchanged. This behavior may stem from the difficulty agents encounter in learning coordinated trading strategies. Another explanation could be that, given the resource scarcity in the applied environment, agents are discouraged from selling the energy as using it to satisfy their own energy needs brings them higher profits. To fully investigate this matter, more research is necessary on environments where trading is proven to be the most optimal decision.

While the effect of changed reward function on the agents' behavior is easily explainable one may ask why the agents achieve much better training outcomes in environments rich in PV generation. One possible explanation is the presence of an energy surplus which provides agents with more flexibility to experiment. In environments where energy production is closely aligned with energy demand any deviation, such as storing extra energy instead of using it immediately, can heavily penalize the agents discouraging them from exploring energy-saving strategies. In this case, failing to meet their energy demand through generation forces the agents to import energy from the retailer, resulting in high costs. In contrast, environments with abundant energy generation grant many more opportunities for learning as agents can afford to experiment with actions such as storing surplus energy that would otherwise be exported to the retailer. It is important to note that we do not reward agents for exporting the energy to the retailer, meaning that the actions of storing and exporting energy are treated identically from the reward perspective.



(a) $\beta_s = 0$



(b) $\beta_s = 10$

Figure 4.17: The distribution of incoming energy sources and outgoing energy destinations, averaged across ten unique households for a realistic scenario with $\eta_{\text{pv}} = 0.5$ and two different values of β_s.

5

Conclusion

The main focus of this thesis was to investigate how Sequential Social Dilemmas (SSDs) emerge and can be mitigated within a practical MARL framework. Due to their growing importance and popularity as a focus for RL applications, Local Energy Communities were chosen to be the setting for this study. We argued that, when certain conditions are present, SSDs within LEC environments can lead to suboptimal training outcomes. To test this, we created a novel RL-driven LEC environment allowing us to face the challenge from the perspective of individual households, each driven by a separate learned policy. PPO was chosen as the guiding RL algorithm due to its proven effectiveness in such settings. The training was performed using a Pecan Street dataset. With this setup in place, a series of experiments was conducted.

We first focused on examining the individual training capabilities of the agents, demonstrating that agents are able to successfully learn to manage their local batteries. However, their efficiency was heavily dependent on their training environment as their training performance notably declined in settings characterized by low availability of solar energy.

Next, we turned our attention to scenarios in which agents interact with one another through the shared community battery. Building on the previous findings, we adapted an existing evaluation procedure to detect the existence of SSDs. The analysis involved training agents under two different sets of environmental conditions - one with upscaled and one with downscaled energy generation levels to simulate varying resource scarcity. Two resulting sets of policies were then matched against one another in resource-scarce conditions. The outcomes confirmed that social dilemmas did occur and their presence was less common in environments with higher energy generation. Moreover, as in the individual case, agents trained in resource-rich environments consistently outperformed those trained under resource-scarce conditions. This trend was also reflected in their level of cooperativeness with agents from resource-rich settings being significantly more likely to contribute to the shared battery than those from resource-scarce environments. One interesting observation that we had was that the agents trained in environments abundant in solar energy did better in resource-poor environments than agents trained directly under those conditions. This suggests that rescaling the training data provides one effective way of pushing agents to choose more cooperative actions in a social dilemma setting.

In the following experiments, we proposed another mitigation strategy to improve the social welfare of the agents. The training procedure was modified by adding a new component to the reward to incorporate community goals into agents' objectives. The new component incentivized contributions to the shared battery and penalized excessive discharging, which allowed us to significantly reduce selfish tendencies, even in resource-scarce settings. Two variants of the component were tested- one that adds a fixed and one that adds a price-based bonus reward for utilizing the common battery in a community-friendly way. While both methods improved the final evaluation results, the price-based factor performed worse than the fixed one.

Finally, we extended our investigation to a more realistic scenario with several households exhibiting diverse characteristics. Our goal was to verify whether our findings hold in more complex environments.

The results revealed that our findings remain valid in the more complex scenarios. Training outcomes are generally better in environments abundant in PV generation. Moreover, we managed to make the agents exhibit more cooperative behavior by applying the reward function modification presented earlier. Interestingly, as shown by the findings, this cooperative behavior was not reflected in more frequent charges of the shared battery but in more coordinated strategies for managing it.

Our first research question focused on methods for identifying and evaluating Sequential Social Dilemmas in MARL settings. We addressed this question through Experiments I and II where we evaluated and matched together policies trained under different resource-availability conditions. Through a thorough analysis of agent profits and interactions with the shared battery, we demonstrated that the dynamics in LECs resemble those of classic game-theoretic models, such as the Prisoner's Dilemma, where agents choose to cooperate by contributing to the communal resource or defect by over-withdrawing from it.

The second research question explored the ways of mitigating these dilemmas to guide the agents' policies toward cooperation and, in turn, lead to growth in social welfare. Results of Experiments I, II, and IV indicate that the training conditions, in our case defined as resource availability, play a pivotal role in this process. Our findings show that by selecting appropriate training data or even transforming the existing datasets (e.g., by upscaling PV generation levels) agents are given more opportunities to learn mutually-beneficial behaviors. Furthermore, in Experiment III, we presented two alternatives to the used objective function which allowed us to achieve satisfying training outcomes, even if the resource availability was very low.

In summary, this thesis made three key contributions. First, we designed and implemented a new agent-centric LEC Python environment which offers a lot of potential for investigating social dilemmas in such settings. Second, with the use of this environment, we demonstrated that SSDs indeed arise and should be accounted for during the training process. As a means to address SSDs in LECs, we proposed two strategies: one that transforms the training data and another that focuses on adjusting the reward function.

As a relatively novel field of research, there are many potential future directions to follow. Future work could involve adding more assets to the investigated environments (e.g. EVs, thermal components) or including objectives other than energy saving. These new objectives could be focused on the reduction of greenhouse gas emissions or maximizing household satisfaction as both cases require a balance between individual and community goals.

In our study, the modifications to the reward function were chosen by hand and applied without the agents' involvement. Another promising avenue for future research would be to fully implement a contracting mechanism resembling this presented in the work by Haupt et al. [11] where agents themselves agreed on alterations to their objectives in order to satisfy community needs.

The custom environment we implemented and used incorporated both trading and shared batteries as a means for agents to interact with one another. However, our SSD mitigation efforts mostly focused on the dynamics related to the shared batteries. Therefore, yet another research direction could follow the scenarios similar to the ones outlined in Subsection 3.2.2 investigating how community-oriented incentives can be injected into the trading mechanisms.

References

[1] Irati Otamendi-Irizar et al. “How can local energy communities promote sustainable development in European cities?” In: *Energy Research & Social Science* 84 (2022), p. 102363. ISSN: 2214-6296. DOI: 10.1016/j.erss.2021.102363.

[2] Edoardo Barabino et al. “Energy Communities: A review on trends, energy system modelling, business models, and optimisation objectives”. In: *Sustainable Energy, Grids and Networks* 36 (2023), p. 101187. ISSN: 2352-4677. DOI: 10.1016/j.segan.2023.101187.

[3] Hussain Kazmi et al. “Towards data-driven energy communities: A review of open-source datasets, models and tools”. In: *Renewable and Sustainable Energy Reviews* 148 (June 2021). DOI: 10.1016/j.rser.2021.111290.

[4] Mona Bielig et al. “Evidence behind the narrative: Critically reviewing the social impact of energy communities in Europe”. In: *Energy Research & Social Science* 94 (2022), p. 102859. ISSN: 2214-6296. DOI: 10.1016/j.erss.2022.102859.

[5] Flora Charbonnier, Thomas Morstyn, and Malcolm D. McCulloch. “Coordination of resources at the edge of the electricity grid: Systematic review and taxonomy”. In: *Applied Energy* 318 (2022), p. 119188. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2022.119188.

[6] Xin Chen et al. *Reinforcement Learning for Decision-Making and Control in Power Systems: Tutorial, Review, and Vision*. Jan. 2021. DOI: 10.48550/arXiv.2102.01168.

[7] Wim Liebrand, David Messick, and Henk Wilke. *Social dilemmas: Theoretical issues and research findings*. Garland Science, 2015.

[8] Robert Axelrod. “Effective Choice in the Prisoner’s Dilemma”. In: *Journal of Conflict Resolution* 24.1 (1980), pp. 3–25. DOI: 10.1177/002200278002400101.

[9] Joel Z. Leibo et al. *Multi-agent Reinforcement Learning in Sequential Social Dilemmas*. 2017. DOI: 10.48550/arXiv.1702.03037. arXiv: 1702.03037 [cs.MA].

[10] Frans Oliehoek and Christopher Amato. “A Concise Introduction to Decentralized POMDPs”. In: (Jan. 2016). DOI: 10.1007/978-3-319-28929-8.

[11] Andreas A. Haupt et al. *Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL*. 2024. arXiv: 2208.10469 [cs.AI].

[12] Andrei Lupu and Doina Precup. “Gifting in Multi-Agent Reinforcement Learning”. In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS ’20. Auckland, New Zealand: International Foundation for Autonomous Agents and Multiagent Systems, 2020, pp. 789–797. ISBN: 9781450375184.

[13] L. P. Kaelbling, M. L. Littman, and A. W. Moore. *Reinforcement Learning: A Survey*. 1996. arXiv: cs/9605103 [cs.AI].

[14] Lucian Busoniu, Robert Babuska, and Bart De Schutter. “A Comprehensive Survey of Multiagent Reinforcement Learning”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38.2 (2008), pp. 156–172. DOI: 10.1109/TSMCC.2007.913919.

[15] Yaodong Yang and Jun Wang. *An Overview of Multi-Agent Reinforcement Learning from Game Theoretical Perspective*. 2021. arXiv: 2011.00583 [cs.MA]. URL: https://arxiv.org/abs/2011.00583.

[16] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. *Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms*. 2021. arXiv: 1911.10635 [cs.LG]. URL: https://arxiv.org/abs/1911.10635.

[17] Michael L. Littman. “Markov games as a framework for multi-agent reinforcement learning”. In: *Machine Learning Proceedings 1994*. Ed. by William W. Cohen and Haym Hirsh. San Francisco (CA): Morgan Kaufmann, 1994, pp. 157–163. ISBN: 978-1-55860-335-6. DOI: 10.1016/B978-1-55860-335-6.50027-1.

[18] Pablo Hernandez-Leal et al. *A Survey of Learning in Multiagent Environments: Dealing with Non-Stationarity*. 2019. arXiv: 1707.09183 [cs.MA].

[19] Jakob N. Foerster et al. *Learning to Communicate with Deep Multi-Agent Reinforcement Learning*. 2016. arXiv: 1605.06676 [cs.AI]. URL: https://arxiv.org/abs/1605.06676.

[20] Chenyu Guo et al. “Real-time optimal energy management of microgrid with uncertainties based on deep reinforcement learning”. In: *Energy* 238 (2022), p. 121873. ISSN: 0360-5442. DOI: 10.1016/j.energy.2021.121873.

[21] Guozhou Zhang et al. “Data-driven optimal energy management for a wind-solar-diesel-battery-reverse osmosis hybrid energy system using a deep reinforcement learning approach”. In: *Energy Conversion and Management* 227 (2021), p. 113608. ISSN: 0196-8904. DOI: 10.1016/j.enconman.2020.113608.

[22] John Schulman et al. *Proximal Policy Optimization Algorithms*. 2017. arXiv: 1707.06347 [cs.LG]. URL: https://arxiv.org/abs/1707.06347.

[23] John Schulman et al. *Trust Region Policy Optimization*. 2017. arXiv: 1502.05477 [cs.LG]. URL: https://arxiv.org/abs/1502.05477.

[24] Chao Yu et al. *The Surprising Effectiveness of PPO in Cooperative, Multi-Agent Games*. 2022. arXiv: 2103.01955 [cs.LG]. URL: https://arxiv.org/abs/2103.01955.

[25] Mancur Olson. *The logic of collective action: public goods and the theory of groups*. en-US. Harvard economic studies 124. Cambridge, Mass.: Harvard Univ. Press, 1965, p. 176. ISBN: 0-674-53751-3.

[26] Robyn M Dawes. “Social dilemmas”. In: *Annual review of psychology* 31.1 (1980), pp. 169–193.

[27] Paul AM Van Lange et al. “The psychology of social dilemmas: A review”. In: *Organizational Behavior and Human Decision Processes* 120.2 (2013), pp. 125–141. DOI: 10.1016/j.obhdp.2012.11.003.

[28] Roger D Congleton. *Solving social dilemmas: Ethics, politics, and prosperity*. Oxford University Press, 2022.

[29] Elinor Ostrom. “Tragedy of the commons”. In: *The new palgrave dictionary of economics* 2 (2008), pp. 1–4.

[30] Brian Skyrms. “The Stag Hunt”. In: *Proceedings and Addresses of the American Philosophical Association* 75.2 (2001), pp. 31–41. DOI: 10.2307/3218711.

[31] Anatol Rapoport and Albert M. Chammah. “The Game of Chicken”. In: *American Behavioral Scientist* 10.3 (1966), pp. 10–28. DOI: 10.1177/000276426601000303.

[32] Michael W Macy and Andreas Flache. “Learning dynamics in social dilemmas”. In: *Proceedings of the National Academy of Sciences* 99.suppl_3 (2002), pp. 7229–7236. DOI: 10.1073/pnas.092080099.

[33] Adam Lerer and Alexander Peysakhovich. *Maintaining cooperation in complex social dilemmas using deep reinforcement learning*. 2018. arXiv: 1707.01068 [cs.AI].

[34] Chao Yu et al. “Emotional Multiagent Reinforcement Learning in Spatial Social Dilemmas”. In: *IEEE Transactions on Neural Networks and Learning Systems* 26.12 (2015), pp. 3083–3096. DOI: 10.1109/TNNLS.2015.2403394.

[35] Nicolas Anastassacos et al. *Cooperation and Reputation Dynamics with Reinforcement Learning*. 2021. arXiv: 2102.07523 [cs.MA].

[36] European Committee of the Regions et al. *Models of local energy ownership and the role of local energy communities in energy transition in Europe*. European Committee of the Regions, 2018. DOI: 10.2863/603673.

[37] Kirsten Jenkins et al. “Energy justice: A conceptual review”. In: *Energy Research & Social Science* 11 (2016), pp. 174–182. ISSN: 2214-6296. DOI: 10.1016/j.erss.2015.10.004.

[38] Kacper Szulecki. “Conceptualizing energy democracy”. In: *Environmental Politics* 27.1 (2018), pp. 21–41. DOI: 10.1080/09644016.2017.1387294.

[39] Andrea M. Feldpausch-Parker, Danielle Endres, and Tarla Rai Peterson. “Editorial: A Research Agenda for Energy Democracy”. In: *Frontiers in Communication* 4 (2019). ISSN: 2297-900X. DOI: 10.3389/fcomm.2019.00053.

[40] Anna Schreuer. “The establishment of citizen power plants in Austria: A process of empowerment?” In: *Energy Research & Social Science* 13 (2016). Energy Transitions in Europe: Emerging Challenges, Innovative Approaches, and Possible Solutions, pp. 126–135. ISSN: 2214-6296. DOI: 10.1016/j.erss.2015.12.003.

[41] Florian Hanke and Jens Lowitzsch. “Empowering Vulnerable Consumers to Join Renewable Energy Communities—Towards an Inclusive Design of the Clean Energy Package”. In: *Energies* 13.7 (2020). ISSN: 1996-1073. DOI: 10.3390/en13071615.

[42] Nan Lin. *Building a network theory of social capital*. Routledge, 2017, pp. 3–28.

[43] Tilman Weckesser et al. “Renewable Energy Communities: Optimal sizing and distribution grid impact of photo-voltaics and battery storage”. In: *Applied Energy* 301 (2021), p. 117408. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2021.117408.

[44] Esa Pursiheimo and Miika Rämä. “Optimal Capacities of Distributed Renewable Heat Supply in a Residential Area Connected to District Heating”. In: *Journal of Sustainable Development of Energy, Water and Environment Systems* N/A (July 2020). DOI: 10.13044/j.sdewes.d8.0328.

[45] Ricardo Faia et al. “An Optimization Model for Energy Community Costs Minimization Considering a Local Electricity Market between Prosumers and Electric Vehicles”. In: *Electronics* 10.2 (2021). ISSN: 2079-9292. DOI: 10.3390/electronics10020129.

[46] Jiang-Wen Xiao et al. “A new energy storage sharing framework with regard to both storage capacity and power capacity”. In: *Applied Energy* 307 (2022), p. 118171. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2021.118171.

[47] Bing Yan et al. “Markovian-based stochastic operation optimization of multiple distributed energy systems with renewables in a local energy community”. In: *Electric Power Systems Research* 186 (2020), p. 106364. ISSN: 0378-7796. DOI: 10.1016/j.epsr.2020.106364.

[48] Pei Huang et al. “Solar-photovoltaic-power-sharing-based design optimization of distributed energy storage systems for performance improvements”. In: *Energy* 222 (2021), p. 119931. ISSN: 0360-5442. DOI: 10.1016/j.energy.2021.119931.

[49] Ezzeddin Bakhtavar et al. “Assessment of renewable energy-based strategies for net-zero energy communities: A planning model using multi-objective goal programming”. In: *Journal of Cleaner Production* 272 (2020), p. 122886. ISSN: 0959-6526. DOI: 10.1016/j.jclepro.2020.122886.

[50] Anthony Maturo et al. “Design and environmental sustainability assessment of energy-independent communities: The case study of a livestock farm in the North of Italy”. In: *Energy Reports* 7 (2021), pp. 8091–8107. ISSN: 2352-4847. DOI: 10.1016/j.egyr.2021.05.080.

[51] Mohammad Sadegh Javadi et al. “Pool trading model within a local energy community considering flexible loads, photovoltaic generation and energy storage systems”. In: *Sustainable Cities and Society* 79 (2022), p. 103747. ISSN: 2210-6707. DOI: 10.1016/j.scs.2022.103747.

[52] Inês F. G. Reis et al. “Assessing the Influence of Different Goals in Energy Communities’ Self-Sufficiency—An Optimized Multiagent Approach”. In: *Energies* 14.4 (2021). ISSN: 1996-1073. DOI: 10.3390/en14040989.

[53] G. Mutani, S. Santantonio, and S. Beltramino. “Indicators and representation tools to measure the technical-economic feasibility of a renewable energy community. The case study of Villar Pellice (Italy)”. In: *International Journal of Sustainable Development and Planning* 16.1 (2021), pp. 1–11. DOI: 10.18280/ijsdp.160101.

[54] Wenzhi Cao et al. “An efficient and economical storage and energy sharing model for multiple multi-energy microgrids”. In: *Energy* 244 (2022), p. 123124. ISSN: 0360-5442. DOI: 10.1016/j.energy.2022.123124.

[55] Zhijian Liu et al. “Multi-scenario analysis and collaborative optimization of a novel distributed energy system coupled with hybrid energy storage for a nearly zero-energy community”. In: *Journal of Energy Storage* 41 (2021), p. 102992. ISSN: 2352-152X. DOI: 10.1016/j.est.2021.102992.

[56] Amir Safdarian et al. “Coalitional Game Theory Based Value Sharing in Energy Communities”. In: *IEEE Access* 9 (2021), pp. 78266–78275. DOI: 10.1109/ACCESS.2021.3081871.

[57] Paul Denholm and Robert M. Margolis. “Evaluating the limits of solar photovoltaics (PV) in traditional electric power systems”. In: *Energy Policy* 35.5 (2007), pp. 2852–2861. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2006.10.014.

[58] Edward Barbour et al. “Community energy storage: A smart choice for the smart grid?” In: *Applied Energy* 212 (2018), pp. 489–497. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2017.12.056.

[59] Mike B. Roberts, Anna Bruce, and Iain MacGill. “Impact of shared battery energy storage systems on photovoltaic self-consumption and electricity bills in apartment buildings”. In: *Applied Energy* 245 (2019), pp. 78–95. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2019.04.001.

[60] Linda Colarullo and Jagruti Thakur. “Second-life EV batteries for stationary storage applications in Local Energy Communities”. In: *Renewable and Sustainable Energy Reviews* 169 (2022), p. 112913. ISSN: 1364-0321. DOI: 10.1016/j.rser.2022.112913.

[61] Boyuan Zhu et al. *Battery-based storage for communities*. 2018. DOI: 10.1049/PBPO130E_ch10.

[62] F Grasso et al. “Peer-to-peer energy exchanges model to optimize the integration of renewable energy sources: The e-cube project”. In: *L’Energia Elettrica* 96 (2019), pp. 0–0. DOI: 10.36156/ENERGIA06_02.

[63] Theresia Perger et al. “PV sharing in local communities: Peer-to-peer trading under consideration of the prosumers’ willingness-to-pay”. In: *Sustainable Cities and Society* 66 (2021), p. 102634. ISSN: 2210-6707. DOI: 10.1016/j.scs.2020.102634.

[64] Valeria Casalicchio et al. “From investment optimization to fair benefit distribution in renewable energy community modelling”. In: *Applied Energy* 310 (2022), p. 118447. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2021.118447.

[65] Sonam Norbu et al. “Modeling economic sharing of joint assets in community energy projects under LV network constraints”. In: *IEEE Access* 9 (2021), pp. 112019–112042. DOI: 10.1109/ACCESS.2021.3103480.

[66] J.M. Schwidtal et al. “Emerging business models in local energy markets: A systematic review of peer-to-peer, community self-consumption, and transactive energy models”. In: *Renewable and Sustainable Energy Reviews* 179 (2023), p. 113273. ISSN: 1364-0321. DOI: 10.1016/j.rser.2023.113273.

[67] Mohsen Khorasany, Yateendra Mishra, and Gerard Ledwich. “Market framework for local energy trading: a review of potential designs and market clearing approaches”. In: *IET Generation, Transmission & Distribution* 12.22 (2018), pp. 5899–5908. DOI: 10.1049/iet-gtd.2018.5309.

[68] Tiago Sousa et al. “Peer-to-peer and community-based markets: A comprehensive review”. In: *Renewable and Sustainable Energy Reviews* 104 (2019), pp. 367–378. ISSN: 1364-0321. DOI: 10.1016/j.rser.2019.01.036.

[69] Thomas Morstyn, Alexander Teytelboym, and Malcolm D. Mcculloch. “Bilateral Contract Networks for Peer-to-Peer Energy Trading”. In: *IEEE Transactions on Smart Grid* 10.2 (2019), pp. 2026–2035. DOI: 10.1109/TSG.2017.2786668.

[70] Etienne Sorin, Lucien Bobo, and Pierre Pinson. “Consensus-Based Approach to Peer-to-Peer Electricity Markets With Product Differentiation”. In: *IEEE Transactions on Power Systems* 34.2 (2019), pp. 994–1004. DOI: 10.1109/TPWRS.2018.2872880.

[71] Esther Mengelkamp et al. “Designing microgrid energy markets: A case study: The Brooklyn Microgrid”. In: *Applied Energy* 210 (2018), pp. 870–880. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2017.06.054.

[72] M. N. Akter, M. A. Mahmud, and Amanullah M. T. Oo. “A hierarchical transactive energy management system for microgrids”. In: *2016 IEEE Power and Energy Society General Meeting (PESGM)*. 2016, pp. 1–5. DOI: 10.1109/PESGM.2016.7741099.

[73] Wayes Tushar et al. “Energy Storage Sharing in Smart Grid: A Modified Auction-Based Approach”. In: *IEEE Transactions on Smart Grid* 7.3 (2016), pp. 1462–1475. DOI: 10.1109/TSG.2015.2512267.

[74] Pol Olivella-Rosell et al. “Day-ahead micro-market design for distributed energy resources”. In: *2016 IEEE International Energy Conference (ENERGYCON)*. 2016, pp. 1–6. DOI: 10.1109/ENERGYCON.2016.7513961.

[75] Tian Liu et al. “Energy management of cooperative microgrids with P2P energy sharing in distribution networks”. In: *2015 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. 2015, pp. 410–415. DOI: 10.1109/SmartGridComm.2015.7436335.

[76] Chao Long et al. “Feasibility of Peer-to-Peer Energy Trading in Low Voltage Electrical Distribution Networks”. In: *Energy Procedia* 105 (2017). 8th International Conference on Applied Energy, ICAE2016, 8-11 October 2016, Beijing, China, pp. 2227–2232. ISSN: 1876-6102. DOI: 10.1016/j.egypro.2017.03.632.

[77] Guro Sæther, Pedro Crespo del Granado, and Salman Zaferanlouei. “Peer-to-peer electricity trading in an industrial site: Value of buildings flexibility on peak load reduction”. In: *Energy and Buildings* 236 (2021), p. 110737. ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2021.110737.

[78] Tilman Weckesser et al. “Renewable Energy Communities: Optimal sizing and distribution grid impact of photo-voltaics and battery storage”. In: *Applied Energy* 301 (2021), p. 117408. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2021.117408.

[79] Giacomo Talluri et al. “Optimal Battery Energy Storage System Scheduling within Renewable Energy Communities”. In: *Energies* 14 (Dec. 2021), p. 8480. DOI: 10.3390/en14248480.

[80] Mircea Simoiu et al. “Sizing and Management of an Energy System for a Metropolitan Station with Storage and Related District Energy Community”. In: *Energies* 14 (Sept. 2021), p. 5997. DOI: 10.3390/en14185997.

[81] Ziqing Zhu et al. “Reinforcement learning in deregulated energy market: A comprehensive review”. In: *Applied Energy* 329 (2023), p. 120212. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2022.120212.

[82] Dawei Qiu et al. “Mean-field multi-agent reinforcement learning for peer-to-peer multi-energy trading”. In: *IEEE Transactions on Power Systems* (2022). DOI: 10.1109/TPWRS.2022.3217922.

[83] A.T.D. Perera and Parameswaran Kamalaruban. “Applications of reinforcement learning in energy systems”. In: *Renewable and Sustainable Energy Reviews* 137 (2021), p. 110618. ISSN: 1364-0321. DOI: 10.1016/j.rser.2020.110618.

[84] Yuankun Liu, Dongxia Zhang, and Hoay Beng Gooi. “Optimization strategy based on deep reinforcement learning for home energy management”. In: *CSEE Journal of Power and Energy Systems* 6.3 (2020), pp. 572–582. DOI: 10.17775/CSEEJPES.2019.02890.

[85] Flora Charbonnier, Thomas Morstyn, and Malcolm D. McCulloch. “Scalable multi-agent reinforcement learning for distributed control of residential energy flexibility”. In: *Applied Energy* 314 (2022), p. 118825. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2022.118825.

[86] Jun Cao et al. “Deep Reinforcement Learning-Based Energy Storage Arbitrage With Accurate Lithium-Ion Battery Degradation Model”. In: *IEEE Transactions on Smart Grid* 11.5 (2020), pp. 4513–4521. DOI: 10.1109/TSG.2020.2986333.

[87] William Valladares et al. “Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm”. In: *Building and Environment* 155 (2019), pp. 105–117. ISSN: 0360-1323. DOI: 10.1016/j.buildenv.2019.03.038.

[88] Lukas Eller, Lydia C. Siafara, and Thilo Sauter. “Adaptive control for building energy management using reinforcement learning”. In: *2018 IEEE International Conference on Industrial Technology (ICIT)*. 2018, pp. 1562–1567. DOI: 10.1109/ICIT.2018.8352414.

[89] Zheng Wen, Daniel O'Neill, and Hamid Maei. "Optimal demand response using device-based reinforcement learning". In: *IEEE Transactions on Smart Grid* 6.5 (2015), pp. 2312–2324. DOI: 10.1109/TSG.2015.2396993.

[90] Tianyi Chen and Shengrong 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: 10.1109/ISGTEurope.2019.8905731.

[91] Seongwoo Lee et al. "Novel Energy Trading System Based on Deep-Reinforcement Learning in Microgrids". In: *Energies* 14 (Sept. 2021), p. 5515. DOI: 10.3390/en14175515.

[92] Jiehui Zheng et al. "Multi-Agent Reinforcement Learning With Privacy Preservation for Continuous Double Auction-Based P2P Energy Trading". In: *IEEE Transactions on Industrial Informatics* 20.4 (2024), pp. 6582–6590. DOI: 10.1109/TII.2023.3348823.

[93] Helder Pereira, Luis Gomes, and Zita Vale. "Peer-to-peer energy trading optimization in energy communities using multi-agent deep reinforcement learning". In: *Energy Informatics* 5.Suppl 4 (2022), p. 44. DOI: 10.1186/s42162-022-00235-2.

[94] Jiatong Wang, Li Li, and Jiangfeng Zhang. "Deep reinforcement learning for energy trading and load scheduling in residential peer-to-peer energy trading market". In: *International Journal of Electrical Power & Energy Systems* 147 (2023), p. 108885. ISSN: 0142-0615. DOI: 10.1016/j.ijepes.2022.108885.

[95] Sania Khaskheli and Amjad Anvari-Moghaddam. "Energy Trading in Local Energy Markets: A Comprehensive Review of Models, Solution Strategies, and Machine Learning Approaches". In: *Applied Sciences* 14.24 (2024). ISSN: 2076-3417. DOI: 10.3390/app142411510.

[96] Pecan Street Inc. *Dataport: The world's largest residential energy dataset*. Accessed: 2024-12-29. 2024. URL: https://www.pecanstreet.org/dataport/.

[97] New York Independent System Operator (NYISO). *NYISO website*. Accessed: 2024-12-29. 2024. URL: https://www.nyiso.com/.

A

PPO Configuration

A.1. Hyperparameter and Model Settings

Table A.1: PPO hyperparameters used in this work.

Parameter	Value	Description
Training Parameters		
lr	5×10^{-5}	Learning rate used by the optimizer.
gamma	0.99	Discount factor for future rewards.
train_batch_size	50000	Total size of the training batch
sgd_minibatch_size	512	Size of each minibatch for SGD updates
num_sgd_iter	10	Number of SGD iterations per training batch .
clip_param	0.15	Clipping parameter for the surrogate loss
lambda	1.0	GAE parameter for the bias-variance tradeoff.
kl_coeff	0.2	Coefficient for the KL-divergence penalty.
kl_target	0.01	Target KL divergence threshold.
clip_actions	False	Flag to clip actions to the action space bounds.
vf_clip_param	10.0	Clipping parameter for the value function loss.
vf_loss_coeff	1.0	Weighting factor for the value function loss.
entropy_coeff	0.0	Coefficient for the entropy bonus.
entropy_coeff_schedule	None	Schedule for adjusting the entropy coefficient.
rollout_fragment_length	auto	Length (in timesteps) of rollout fragments.
grad_clip	None	Maximum gradient norm.
use_gae	True	Flag for Generalized Advantage Estimation (GAE).
use_critic	True	Whether to use critic.
batch_mode	truncate_episodes	Method for batching episodes.
num_workers	15	Number of parallel rollout workers.
Model Parameters		
use_lstm	False	Flag to use an LSTM-based recurrent network.
max_seq_len	20	Maximum sequence length for LSTM training.
vf_share_layers	False	Should value and policy networks share layers.
fcnet_hiddens	[256, 256]	Fully connected network hidden layers
fcnet_activation	relu	Activation function for the FC network