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

Daniel C. May[a], Matthew Taylor[a], Petr Musilek[a,c,*]

[a] *Electrical and Computer Engineering, University of Alberta, Edmonton, T6G 2R3, AB, Canada*
[b] *Computing Science, University of Alberta, Edmonton, T6G 2R3, AB, Canada*
[c] *Applied Cybernetics, University of Hradec Králové, 500 03, Hradec Králové, Czech Republic*

Abstract

As the energy landscape evolves toward sustainability, the accelerating integration of distributed energy resources poses challenges to the operability and reliability of the electricity grid. One significant aspect of this issue is the notable increase in net load variability at the grid edge.

Transactive energy, implemented through local energy markets, has recently garnered attention as a promising solution to address the grid challenges in the form of decentralized, indirect demand response on a community level. Given the nature of these challenges, model-free control approaches, such as deep reinforcement learning, show promise for the decentralized automation of participation within this context. Existing studies at the intersection of transactive energy and model-free control primarily focus on socioeconomic and self-consumption metrics, overlooking the crucial goal of reducing community-level net load variability.

This study addresses this gap by training a set of deep reinforcement learning agents to automate end-user participation in ALEX, an economy-driven local energy market. In this setting, agents do not share information and only prioritize individual bill optimization. The study unveils a clear correlation between bill reduction and reduced net load variability in this setup. The impact on net load variability is assessed over various time horizons using metrics such as ramping rate, daily and monthly load factor, as well as daily average and total peak export and import on an open-source dataset. Agents are then benchmarked against several baselines, with their performance levels showing promising results, approaching those of a near-optimal dynamic programming benchmark.

*Corresponding author
Email address: pmusilek@ualberta.ca (Petr Musilek)

Keywords: Reinforcement Learning, Deep Reinforcement Learning, Distributed Energy Resources, Local Energy Markets, Demand Response, Distributed Energy Resource Management, Transactive Energy

1. Introduction

Progress towards sustainable energy utilization is crucial for addressing climate change. In this context, the convergence of technological advances and lagging regulatory frameworks has precipitated the rapid adoption of distributed energy resources (DERs), reshaping the dynamics of the grid edge where electricity end-users reside [1]. Consequently, the variability of the net load at the grid edge is rapidly increasing. The term variability encompasses the composite effects of intermittency and other net load volatilities, such as those caused by electric vehicle charging. This marked increase amplifies the challenges associated with ensuring the reliability and efficiency of grid operations [2, 3]. This drives the transition to the Smart Grid, which operates in a decentralized and autonomous manner to maintain and possibly enhance the operability of the electricity grid.

To address these challenges, the research community has been actively exploring demand response (DR) methodologies. Broadly speaking, DR techniques leverage various signals to modulate end-user load demand, supporting electrical grid efficiency and reliability. These signals encompass both direct control commands to assets and incentive mechanisms intended to influence end-user behavior, thus delineating between direct and indirect DR. Notably, the key hurdles in indirect DR lie in aligning the interests of grid stakeholders and electricity end-users through appropriate incentive structures and subsequently ensuring sufficient participation to achieve the desired effect [4, 5, 6].

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

In response to the challenges faced by these scheduling-based methods, transactive energy (TE) has emerged as a compelling alternative. TE, defined as "the use of a combination of economic and control techniques to improve grid reliability and efficiency" by the GridWise Architecture Council [7], aligns well with the Smart Grid ethos, emphasizing the market as

a decentralized delivery mechanism for incentive signals [4].

Recent literature has highlighted the concept of Local Energy Markets (LEMs) as a viable path to implement TE within geographically constrained communities at the grid edge. Mengelkamp et al. define LEM as "a geographically distinct and socially close community of residential prosumers and consumers who can trade locally produced electricity within their community. For this, all actors must have access to a local market platform on which (buy) bids and (ask) offers for local electricity are matched" [8]. LEMs allow for the delivery of real-time incentive signals to electricity end-users, providing the necessary granularity and immediacy within a decentralizable framework.

The surveys of completed DR pilot studies confirm that automation is necessary to facilitate sufficient levels of participation [4, 5, 6]. While MPC is entrenched in the general DR literature for automation, model-free approaches such as deep reinforcement learning (DRL) present a promising paradigm better suited to tackle the challenges faced at the grid edge. Initially inspired by high-level performance showcases of DRL in games [9, 10, 11], this notion is reinforced by the success of DRL in fields like robotics [12] and process control [13]. Moreover, it is supported by a growing body of research applying DRL to the electricity grid [14, 15, 16].

Within this context, recent studies have explored automating end-user participation and DER management in LEMs [17, 18, 19, 20, 21, 22, 23], predominantly through agents trained to optimize end-user bills via load-shifting capacities. Some studies demonstrate the reduction of net community energy consumption [19, 21], while others investigate the provision of flexibility services [18]. However, to the best of the authors' knowledge, there are no other studies demonstrating the reduction of community-level load variability through the automation of LEMs using DRL.

Such a conclusive demonstration is not trivial. Despite the intention of LEMs to align the interests of end-users with the objectives of grid stakeholders, it is crucial to recognize that incentivized behavior may not automatically translate into reduced variability or enhanced power quality at the local level [24, 25]. Similarly, the intricate interplay between LEM design and participant automation may yield unforeseen outcomes [26], a phenomenon commonly observed when automating complex systems using DRL [27].

This article addresses this research gap by training independent agents to automate end-user participation in LEMs and the utilization of DERs. The study demonstrates an emergent reduction in community net load variability even when agents solely prioritize individual bill

optimization. To enhance benchmarking and future comparability, performance evaluation is conducted on an open-source dataset, and the agent's performance is compared to several baselines. The trained DRL agents perform close to the near-optimal benchmark without information sharing or access to future information.

Subsequent sections of this article delve into related work and background in Section 2, methodology for training DRL agents, evaluation and benchmarking procedures in Section 3, a comprehensive discussion of simulation results in Section 4, and conclude with a brief summary and avenues for future research in Section 5.

2. Related Work and Background

Subsection 2.1 briefly reviews related literature and establishes a notable research gap: the lack of a well-benchmarked demonstration of variability reductions within an economy-driven LEM, emerging from selfish end-user bill minimization that DRL agents automate. Subsection 2.2 introduces the LEM design that forms the foundation of this study. Subsection 2.3 overviews reinforcement learning and proximal policy optimization, the base DRL algorithm employed within this article.

2.1. Related Literature

The application of DRL in DR, and for the electricity grid in general, has garnered significant attention in recent years [14, 16, 15]. Studies exploring the distributed coordination of DERs through DR mechanisms outside of LEM, such as those by Chung et al. [28], Zhang et al. [29], and Nweye et al. [30], tend to optimize for composite rewards and incorporate community-level metrics related to grid stability or variability, following a direct optimization approach.

Concurrently, there has been a surge in literature investigating LEMs. Mengelkamp et al. [8], Capper et al. [31], and Tushar et al. [6] provide comprehensive insights into the evolving LEM ecosystem. In general, this field tends to focus on the socioeconomic performance of the proposed system, while DR aspects are only narrowly discussed, and performance benchmarking tends to be restricted.

For instance, Liu et al. [32] propose a LEM-like mechanism, using pricing based on the supply-demand ratio to coordinate energy flow between microgrids, leveraging MPC for automation. Similarly, Lezama et al. [33] explore LEMs from a grid integration perspective, focusing on socioeconomic performance. Ghorani et al. [34] develop bidding models for risk-neutral and risk-averse LEM agents, evaluating their socioeconomic efficacy under various

market designs. Meanwhile, Mengelkamp et al. [26] investigate different market designs using heuristic agents, focusing on socioeconomic metrics. A burgeoning body of research emphasizes the automation of LEM participation through DRL. Xu et al. [19] employ a MARL Q-learning algorithm to automate participation in a LEM that communicates a pricing schedule based on a supply and demand forecast. Zhou et al. [22] propose an economy-driven LEM pricing mechanism, optimizing participant bidding via a combination of Q-learning and fuzzy logic. Similarly, Zang et al. [23] train end-user agents to interact with community-level batteries within LEMs.

As Mengelkamp et al. [26] highlight, the integration of LEMs and automated participation presents complex challenges and potentially unforeseen consequences due to the emergent, intricate system dynamics. Investigations by Kiedanski et al. [24] and Papadaskalopoulose et al. [25] demonstrate that increases in socioeconomic performance in such settings may not directly translate to improved grid performance in terms of reducing variability or improving power quality.

To address this issue, some studies incorporate electricity grid performance metrics into the LEM's pricing mechanism or the agent's reward function, diverging from the original purely economic focus of LEMs and adopting a direct optimization approach. For example, Chen et al.[21] investigate microgrid trading in the context of LEMs, employing a reward function with explicit constraints. Their findings demonstrate that this approach increases self-sufficiency compared to expert-designed heuristics and random action agents in benchmarking experiments. Similarly, Ye et al. [18] explore the use of LEMs to provide flexibility services. Their contribution stands out by benchmarking against a near-optimal MPC baseline, establishing a reasonable upper performance limit. However, even such contributions do not evaluate their agents' performance on variability-related metrics for which the agents do not explicitly optimize.

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

closing this research gap.

2.2. Autonomous Local Energy eXchange

The Autonomous Local Energy eXchange (ALEX), initially proposed by Zhang et al. [35], serves as an LEM for a community denoted as B, where individual buildings $b \in B$ participate in energy trading facilitated by a round-based, futures-blind double auction settlement mechanism. In the context of a round-based futures market, trading occurs in predefined time intervals. A futures market accepts bids and asks for a future settlement timestep t_{settle}, to be submitted at the current time step t_{now}. Settlements are then executed at a subsequent time step t_{deliver}, with $t_{\text{deliver}} > t_{\text{settle}} > t_{\text{now}}$. In such a blind double auction market, each building interacts with the market without awareness of other buildings' activities.

In ALEX, market participants do not share information and instead selfishly optimize their individual electricity bills. The building's electricity bill consists of two main components: the market bill and the grid bill. The market bill includes the cumulative settlement cost, determined by pairing each settlement price with its corresponding quantity for the building. In contrast, the grid bill covers the residual amount required to meet the household's energy demand, billed at the prevailing grid rate selling or buying price, depending on the current net-billing scenario. A profitability margin between the grid rate selling and buying prices serves as an incentive for LEM utilization. This means that any exchange over the LEM presents a favorable scenario. Achieving this could involve leveraging tracked greenhouse gas emission savings or partial fee offsets [31, 36].

Zhang et al. [35] delve into the essential properties required for ALEX's settlement mechanism to incentivize RL agents to learn pricing in correlation with the settlement timestep t_{settle} supply and demand ratios. Formulating ALEX as a mixed-form stochastic game suggests the existence of at least one Nash equilibrium. This insight facilitates the identification of a market mechanism possessing the desired properties through experiments that employ tabular Q-learning bandits under varied but fixed supply and demand ratios. Subsequent experiments deduce a market price function based on the current supply and demand ratio. A follow-up study by Zhang and Musilek [37] investigates a system incorporating a communal battery energy storage system (BESS) controlled by an expert-designed heuristic. The study demonstrates efficacy in avoiding violations of voltage-frequency constraints on a test circuit.

May and Musilek [38] further examines ALEX as a DR system. The authors simulate a group of near-optimal, rational actors on ALEX using an iterative best-response and dynamic

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

This study aims to extend these results by training a set of DRL agents on the equivalent task without access to perfect forecasts, yet achievening a comparable level of variability reduction. This would effectively address the research gap identified in subsection 2.1.

2.3. Reinforcement Learning

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

As illustrated in Figure 1, RL focuses on optimizing the behavior of an agent that interacts with the environment through actions and subsequently receives observations and rewards.

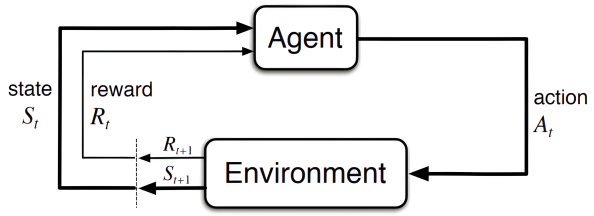

Figure 1: Agent to environment interaction diagram, taken from Sutton & Barto [39].

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

MDPs within the context of RL are typically time-discrete, allowing the notation of the time-step t to represent a specific point in the interaction trajectory between the agent and

the environment. This trajectory starts at $t = 0$ and concludes at $t = T$. The return G signifies the cumulative, discounted future reward,

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

which facilitates the definition of state value

$$V_\pi(s_t) = \mathbf{E} G_t \forall \pi, \tag{2}$$

and state-action value

$$Q_\pi(s_t, a_t) = \mathbf{E} G_t \forall \pi, \tag{3}$$

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

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

$$L(\theta) = \mathbb{E}\left[\log \pi_\theta(a_t, s_t) V_t\right], \tag{4}$$

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

$$L(\theta) = \mathbb{E}\left[\log \pi_\theta(a_t, s_t) A_t\right], A_t = V_t - V_\theta(s_t). \tag{5}$$

Deep Reinforcement Learning (DRL), an amalgamation of RL and deep neural networks, has gained traction for its ability to solve complex MDPs in a generalized manner [10, 11, 9]. DRL methods leverage replay buffers to store agent-environment interactions, facilitating multiple mini-batch stochastic gradient descent epochs. This necessitates the differentiation

between the parameter set used to collect samples into the replay buffer θ_{old} and the new parameters θ, which emerge as a result of gradient updates.

Particularly noteworthy within the dynamic landscape of DRL is Proximal Policy Optimization (PPO), introduced by Schulman et al. [40]. In contrast to naive actor-critic approaches, PPO employs a clipped surrogate objective based on the probability ratio $r(\theta)$. This ratio compares the probabilities of the new policy π_θ and the old policy $\pi_{\theta_{old}}$, aiming to mitigate policy drift and ensure the reliability of data collected into the replay buffer. PPO's actor loss clips the magnitude of the policy ratio $r(\theta)$ within a tolerance parameter ϵ

$$L(\theta) = \mathbb{E}\left[\min\left(r(\theta)A_t, clip(r(\theta), 1-\epsilon, 1+\epsilon)A_t\right)\right]. \tag{6}$$

In addition, most Proximal Policy Optimization (PPO) implementations incorporate generalized advantage estimation, a technique proposed by Schulman et al. [41], to reduce the variance of the advantage A.

3. Methodology and Evaluation

This study aims to extend previous contributions [35, 38] by training DRL agents to autonomously participate in ALEX. The expectation is that these agents will demonstrate a level of emergent community-level variability reduction that is comparable to the near-optimal search method described by May and Musilek [38], but without relying on a perfect forecast. Such a showcase of variability reduction within a DRL-driven LEM context would address the significant research gap outlined in Subsection 2.1.

To achieve this goal, this section formulates ALEX environment as a Markov Decision Process (MDP) in Subsection 3.1, outlines the DRL algorithm employed for training the agents in Subsection 3.2, and elucidates the experimental design in Subsection 3.3. The latter also includes details on evaluation performance metrics and baselines.

3.1. Autonomous Local Energy eXchange as Markov Decision Process

The formulation of ALEX as an MDP involves defining the agent's observations O, actions a, rewards r, and policy π. In comparison to the initial formulation [38], the approach outlined here incorporates specific adaptations tailored to the nature of ALEX as a futures market. This is crucial, given the constraint that the DRL agents should not rely on future information. Additionally, the formulation accommodates continuous observation and action spaces for the DRL agents.

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

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

$$\begin{aligned} O_t^b := (\ & sin(t)_{year}, cos(t)_{year}, sin(t)_{day}, cos(t)_{day}, \\ & E_t^b, SoC_t^b, p_{t_{\text{last settled}}}, q^{bid}_{t_{\text{last settled}}}, q^{ask}_{t_{\text{last settled}}}\ ). \end{aligned} \tag{7}$$

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

In contrast to the action space proposed by Zhang et al. [35], the action space A^b for an agent at timestep t exclusively includes the continuous battery action, scheduled for the future settlement time step $a_{BESS,t_{\text{settle}}}$. This action is constrained by the battery's charge and discharge rates. The determination of bid and ask quantities at settlement time t_{settle} relies on the residual net load, while bid and ask market conditions dictate prices following the round's closure, guided by the price curve defined by Zhang et al. [35].

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

$$a^b_{BESS,t_{\text{settle}}} := a^{\pi_0}_{BESS,t_{\text{settle}}} + a^{\pi_\theta}_{BESS,t_{\text{settle}}}. \tag{8}$$

This action and agent policy definition offers several distinct advantages, significantly

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

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

$$r_t^b := bill_{t_{\text{settle}}}^b - bill_{t_{\text{settle}}}^{b,\pi_0} \tag{9}$$

3.2. Shared Experience Recurrent Proximal Policy Optimization

The agents in this study undergo training as independent agents with shared experience [42]. Although each agent acts autonomously and solely accesses building-level information, they aggregate trajectories into a shared replay buffer. During trajectory collection, the actors function as independent copies of the same actor and critic neural network, which is updated from the shared replay buffer. This maintains full independence between agents during rollout but promotes faster convergence. Christianos et al. [42] demonstrated the efficacy of this approach in enhancing performance within complex multi-agent environments when compared to a fully independent learning setup. Observations undergo standardization and mean-shifting, while rewards are solely standardized, following best practices proposed by Schulman et al. [43].

The remaining portion of this section details modifications to the underlying PPO algorithm. A recurrent PPO [44], using a Long Short-Term Memory (LSTM) [45] hidden layer for both the actor and the critic, is enhanced with recurrent burn-in and initialization, proposed by Kapturowski et al.[46]. Drawing motivation from the findings of Andrychowicz et al. [47], after processing a replay buffer, the new weights θ are used to recalculate the hidden states of the agent LSTM based on the entire trajectory experienced during the current episode. Both enhancements address the risk of stale or drifted state representations, enhancing the agent's capacity to develop meaningful state representations and a long-term context. Informed by Ilyas et al. [48] and with the goal of convergence towards a Nash equilibrium, the learning rate is annealed throughout the training. Instead of setting the value for the terminal transition at T to 0, this study takes it from the critic's value prediction, with preliminary tests indicating an accelerated convergence of the critic to a higher explained

variance. Furthermore, instead of naively imposing action space boundaries by clipping the Gaussian distribution, the algorithm used in this study employs a squashed Gaussian distribution followed by renormalization, as popularized by soft actor-critic algorithms [49].

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

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

3.3. Experimental Design

The DRL agents are trained and evaluated on the CityLearn2022 dataset [50]. The open-source nature of this dataset enables subsequent studies to directly benchmark against this contribution across a diverse range of DR applications. This dataset provides a year of hourly data for 17 smart community buildings, featuring time series of energy demand, photovoltaic generation, and BESS performance characteristics. For each building, one independently acting agent is trained as outlined in Subsection 3.2. Therefore, one episode is defined as a full trajectory over the dataset and lasts 8760 steps, while one run fully trains such a set of 17 agents. For evaluation, the parameter set θ with the best episodic communal return G^B is selected from a run, assuming that this snapshot represents the best-performing equilibrium between agents. This snapshot is updated throughout training when a new best communal return is achieved. From a set of 5 runs, the median performing run is selected for benchmarking purposes.

To assess agent performance, we employ a set of metrics from May and Musilek [38]. All performance metrics in this study are functions of the community net load E^B, defined as the

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

- The average daily imported energy

$$\overline{E}_{d,+} = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(\sum_{t \in d} \max(E^B(t), 0) \right) \tag{10}$$

- The average exported energy

$$\overline{E}_{d,-} = \frac{-1}{n_d} \sum_{d=0}^{n_d} \left(\sum_{t \in d} \min(E^B(t), 0) \right) \tag{11}$$

- The average daily peak

$$\overline{P}_{d,+} = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(\max_{t \in d} E^B(t) \right) \tag{12}$$

- The average daily valley

$$\overline{P}_{d,-} = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(\min_{t \in d} E^B(t) \right) \tag{13}$$

- The absolute maximum peak

$$P_+ = \max_{t=0}^{T} E^B(t) \tag{14}$$

- The absolute minimum valley

$$P_- = \min_{t=0}^{T} E^B(t) \tag{15}$$

- The average daily ramping rate

$$\overline{R}_d = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(\sum_{t \in d} |\nabla E^B(t)| \right) \tag{16}$$

- The daily load factor complement

$$1 - L_d = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(1 - \frac{\text{mean}_{t \in d} E^B(t)}{\max_{t \in d} E^B(t)} \right) \tag{17}$$

- The monthly load factor complement

$$1 - L_m = \frac{1}{n_m} \sum_{m=0}^{n_m} \left(1 - \frac{\text{mean}_{t \in m} E^B(t)}{\max_{t \in m} E^B(t)} \right). \tag{18}$$

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

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

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

- **IndividualDERMS**: In this benchmark, each building in the community operates under a net billing strategy. Buildings prioritize self-sufficiency by smoothing building-level peaks and valleys while minimizing the ramping rate [38]. This benchmark serves as a reasonable performance baseline, resembling a well-tuned heuristic system commonly found in current DR applications. Importantly, unlike the proposed DRL agents, this benchmark has access to a perfect forecast.

- **ALEX DP**: This benchmark represents a near-optimal policy within a discretized version of ALEX's MDP. It is determined using a dynamic programming search method based on iterative best response and value iteration [38]. Importantly, unlike the proposed DRL agents, this benchmark has access to a perfect forecast.

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

4. Results and Discussion

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

The training methodology of the agents focuses on their relative improvement compared to the self-sufficiency maximizing policy π^0, as outlined in Section 3.2. Convergence behaviors are visually depicted in Figure 2, highlighting the average building bill savings of ALEX RL across episodes, benchmarked against ALEX DP. The shaded area represents the variance between runs.

As evident from Figure 2, ALEX RL manages to achieve bill savings that slightly exceed those of ALEX DP. It is crucial to note that ALEX DP performs its search for one day ahead, while ALEX RL is not constrained in the duration of its load shifting. These results indicate that, for the CityLearn 2022 community, there is ample opportunity to shift load over several days.

To strengthen the correlation between achieved bill savings and evaluation metrics, Figure 3 tracks the performance of the median-performing run of ALEX RL in terms of performance evaluation metrics throughout training. A discernible downward trend is evident for all performance metrics, signifying a clear correlation between selfish bill minimization and the selected set of performance metrics.

Qunatitative analysis, summarized in Table 1, consistently supports correlations between performance metrics and participant bill savings. These findings affirm that training DRL

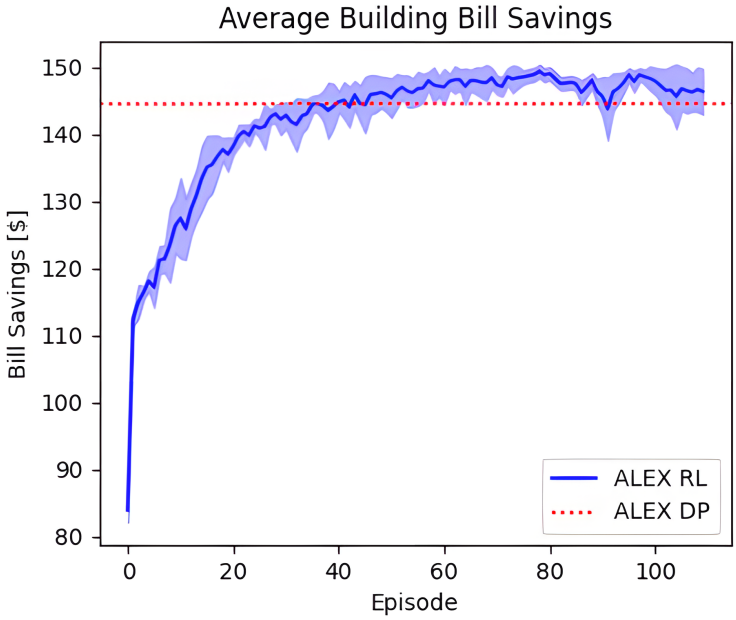

Figure 2: Average participant bill savings comparison between ALEX RL (blue), ALEX DP (red). Shaded areas depict variance bands between a set of 5 ALEX RL runs, trained over 117 episodes.

agents within ALEX incentivize behavior conducive to the emergent suppression of variability in community net load.

These results strongly suggest that the observed correlations between performance metrics and return are consistent across runs. Furthermore, the observed maximum return correlations are consistently higher than the episodic equivalent. Considering ALEX's nature as a mixed-form stochastic game, this outcome is not necessarily surprising and might result from the convergence path towards a Nash equilibrium. This implies that episodes with higher returns tend to be episodes where the agent policies are closer to a joint best response scenario.

The performance of the median performing set of DRL agents is compared to the proposed benchmarks in Table 2. As a result of significantly enhancing the utilization of locally available energy, both the average daily import ($\overline{E_{d,+}}$) and export ($\overline{E_{d,-}}$) decline by 21.9% and 84.4%, respectively. Additionally, emergent peak-shaving behavior leads to a lowering of the average daily peak ($\overline{P_{d,+}}$) and valley ($\overline{P_{d,-}}$) by 27.0% and 71.1%, respectively, while the maximum peak (P_+) and minimum valley (P_-) also shrink by 16.0% and 27.0%, respectively. This behavior also results in the smoothing of moment-to-moment community net-load demand, leading to a 26% decrease in the ramping rate ($\overline{R_d}$) and a mitigation of the overall community net-load swing, which reduces the daily load factor ($1 - L_d$) and monthly

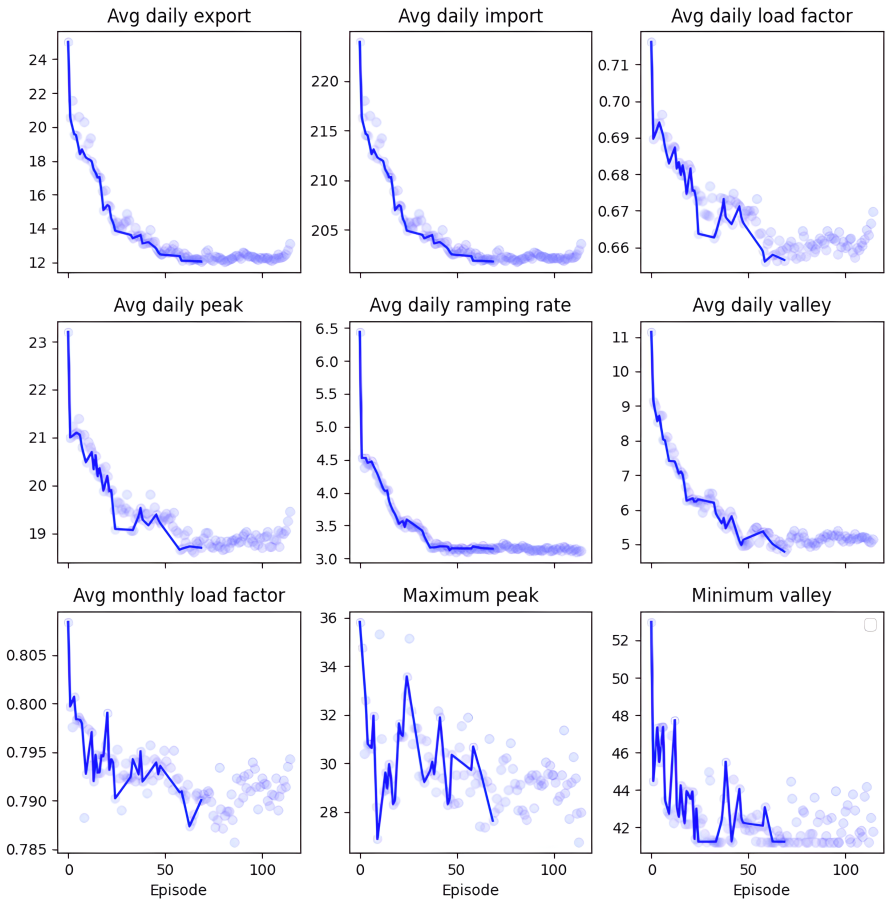

Figure 3: Performance of recorded community-level metrics per episode throughout training. The opaque scattered data points represent singular episode equivalents, while the blue line depicts the metric performance of the most recent highest return achieved.

load factor $(1 - L_m)$ by 11.0% and 3.6%, respectively. In summary, ALEX RL significantly mitigates the effects of community-level variability across all measured metrics.

Further comparative analysis demonstrates the cumulative outperformance of the DRL agents against IndividualDERMS and partial outperformance against ALEX DP. Notably, the ramping rate ($\overline{R_d}$) emerges as a sub-performant metric for ALEX RL compared to IndividualDERMS and ALEX DP. Additionally, it is noteworthy that the average daily valley metric ($\overline{P_{d,-}}$) for ALEX RL is significantly higher than ALEX DP, which is somewhat unexpected. While IndividualDERMS and ALEX DP search over a perfect forecast, ALEX RL does not have access to future information and must internally perform some degree of

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

Table 1: Pearson's correlations between the metrics and achieved bill savings; the rightmost column correlates Maximum Return episodes and their respective metric performance, while the middle column correlates episodic return and the respective episodic metric performance; the bracketed number denotes the average correlation over 5 training runs, whereas the non-bracketed number denotes the correlation of the run achieving the median return.

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

Table 2: Summarized metrics for full simulation on CityLearn2022 data set [50] for NoDERMS, IndividualDERMS and ALEX DP and ALEX DRL scenarios. Values for the NoDERMs, IndividualDERMS and ALEX DP are taken out of May et al. [38]. Best values are typeset in bold.

participant net load modeling. As the most short-term volatility-focused metric, the ramping rate ($\overline{R_d}$) is also most sensitive to such misadjustments. The relative disparity in average daily valley ($\overline{P_{d,-}}$) between ALEX RL and ALEX DP may result from a strategic tradeoff, where it is economically safer for the DRL agents to err on the side of selling to the grid than buying from it in the face of an imperfect model. Such a scenario could occur when the market receives significantly more bids than asks in terms of quantity, as the remaining residual load will be settled according to a net-billing scenario.

These results further suggest that ALEX RL compensates for its lack of perfect internal modeling by leveraging its capability to load shift over a longer duration than ALEX DP, resulting in a further decrease in the maximum peak (P_+) and minimum valley (P_-). Therefore, ALEX RL's relative outperformance in terms of bill savings does not necessarily translate to a strict outperformance of ALEX DP in terms of evaluation metrics. Overall,

ALEX RL's performance closely aligns with ALEX DP, indicating similar levels of emergent, community-level coordination of DERs. The collective results compellingly demonstrate emergent, community-level variability reduction facilitated by automated participation via DRL agents within a LEM, effectively closing the identified research gap.

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

5. Conclusion

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

The rapid proliferation of DERs at the grid edge has led to a significant increase in variability and variance in community net load, posing challenges to electricity grid operability. In response, there has been a growing interest in TE-based DR, facilitated by community LEMs, as a viable solution to align the interests of electricity end-users and grid stakeholders [8, 31, 36]. At the same time, insights from DR system pilots highlight the necessity for automation to ensure robust participation across DR initiatives [5, 4]. In response to the decentralized and distributed nature of this challenge, model-free control approaches, particularly DRL, have emerged as promising candidates [14], fueling the interest in studies investigating the automation of participation in LEMs via DRL methods [19, 22, 23, 21, 18, 17]. While prior research has predominantly focused on socioeconomic metrics and community net load consumption, there remains a gap in demonstrating a clear reduction in variability or variance.

This article addresses the research gap by utilizing a shared experience [42], recurrent PPO [44] algorithm with several modifications [48, 46, 47] to train a set of DRL agents within the context of ALEX, an economy-driven LEM where participants aim to selfishly minimize bills without information sharing [35].The trained DRL agents are compared against benchmark approaches, including a building-level DR strategy and a near-optimal dynamic programming-based solution [38]. Performance is evaluated using a set of metrics capturing net load variance across multiple time horizons, encompassing ramping rate, daily and monthly load factor, peak and average daily import and export. The experiments reveal a clear correlation between relative bill reduction and improvements in the investigated metrics. The trained DRL agents demonstrate promising performance, nearing and, in some

instances, surpassing the benchmarks set by the near-optimal approach, while consistently outperforming the building-level DR strategy.

Future research directions should focus on designing more sophisticated DRL algorithms explicitly tailored to the mixed-form stochastic game nature of LEMs like ALEX. The goal is to establish a clearer performance ceiling for such solutions. Additionally, extending this investigation to diverse LEM designs could offer insights into the factors influencing the efficacy of incentivizing desired behaviors within these systems [24, 25].

Appendix A. Hyperparameters

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

The algorithm employed in this study is rooted in the publicly accessible Recurrent PPO implementation from Stable Baselines3 (SB3)[44]. The hyperparameter values that deviate from SB3's recurrent PPO default settings are as follows:

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

The algorithm adaptations, design, and hyperparameter choices underwent testing across increasingly complex versions of the experiments discussed in the main body of this article until the performance detailed in the discussion section was achieved. The testing progression began with artificial load profiles, aiming to optimize net billing, then advanced to optimizing

net billing on the City Learn dataset for a singular month, then the full year, and finally transitioned to the target application.

The advantage of this iterative process lies in the clearly defined optimal returns for the test scenarios. Recurrent PPO variants seem to vary across implementations, as the exact nature of making PPO recurrent is up to interpretation. We refer to Pleines et al. [51] for an investigation into the characteristics and sensitivities of recurrent PPO. Tests commenced with the default SB3 recurrent PPO in a shared experience replay setting [42], followed by the implementation of R2D2 [46], then state recalculation [47], and finally incorporating a learning rate schedule [48]. Each implementation underwent testing over a small range of hyperparameters for three runs each to ensure consistency, leading to the crystallization of the hyperparameter set used in this study.

The quality of a run was primarily evaluated based on its achieved return, supplemented by the investigation of various RL agent performance metrics. These metrics, inspired by those discussed in the SB3 documentation and Huang et al.'s insightful blog [52], encompassed explained variance, KL-divergence, and entropy loss curves. Even if an algorithm change did not directly impact the agent's average return, it was considered an improvement if, for example, it led to higher explained variance and thereby a stronger critic.

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

Acknowledgment

This research has been supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada grant RGPIN-2017-05866 and by the NSERC/Alberta Innovates grant ALLRP 561116-20.

References

[1] I. R. E. Agency, Global energy transformation: A roadmap to 2050 (2019 edition), Tech. rep., International Renewable Energy Agency (2019).
URL https://www.irena.org/publications/2019/Apr/Global-energy-transformation-A-roadmap-to-2050-2019Edition.

[2] K. Kok, S. Widergren, A society of devices: Integrating intelligent distributed resources with transactive energy, IEEE Power and Energy Magazine 14 (3) (2016) 34–45. doi:10.1109/MPE.2016.2524962.

[3] A. O'Connell, J. Taylor, J. Smith, L. Rogers, Distributed energy resources takes center stage: A renewed spotlight on the distribution planning process, IEEE Power and Energy Magazine 16 (6) (2018) 42–51. doi:10.1109/MPE.2018.2862439.

[4] S. Chen, C.-C. Liu, From demand response to transactive energy: state of the art, Journal of Modern Power Systems and Clean Energy 5 (1) (2017) 10–19. doi:10.1007/s40565-016-0256-x.

[5] F. E. R. Commission, Assessment of demand response and advanced metering, Tech. rep., Federal Energy Regulatory Commission (2015).
URL https://www.ourenergypolicy.org/wp-content/uploads/2015/12/demand-response.pdf

[6] W. Tushar, C. Yuen, T. K. Saha, T. Morstyn, A. C. Chapman, M. J. E. Alam, S. Hanif, H. V. Poor, Peer-to-peer energy systems for connected communities: A review of recent advances and emerging challenges, Applied Energy 282 (2021) 116131. doi:https://doi.org/10.1016/j.apenergy.2020.116131.
URL https://www.sciencedirect.com/science/article/pii/S0306261920315464

[7] R. B. Melton, Gridwise transactive energy framework, Tech. rep., The GridWise Architecture Council (2019).

[8] E. Mengelkamp, J. Diesing, C. Weinhardt, Tracing local energy markets: A literature review:, it - Information Technology 61 (2-3) (2019) 101–110. doi:doi:10.1515/itit-2019-0016.
URL https://doi.org/10.1515/itit-2019-0016

[9] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529–533. doi:https://doi.org/10.1038/nature14236.

[10] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, D. Hassabis, A general

reinforcement learning algorithm that masters chess, shogi, and go through self-play, Science 362 (6419) (2018) 1140–1144. arXiv:https://www.science.org/doi/pdf/10.1126/science.aar6404, doi:https://doi.org/10.1126/science.aar6404.
URL https://www.science.org/doi/abs/10.1126/science.aar6404

[11] OpenAI, :, C. Berner, G. Brockman, B. Chan, V. Cheung, P. Debiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. Pachocki, M. Petrov, H. P. d. O. Pinto, J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang, F. Wolski, S. Zhang, Dota2 with large scale deep reinforcement learning, arXiv preprint arXiv:1912.06680 (2019). arXiv:1912.06680, doi:https://doi.org/10.48550/arXiv.1912.06680.

[12] B. Singh, R. Kumar, V. P. Singh, Reinforcement learning in robotic applications: a comprehensive survey, Artificial Intelligence Review (2022) 1–46doi:https://doi.org/10.1007/s10462-021-09997-9.

[13] J. Shin, T. A. Badgwell, K.-H. Liu, J. H. Lee, Reinforcement learning – overview of recent progress and implications for process control, Computers and Chemical Engineering 127 (2019) 282–294. doi:https://doi.org/10.1016/j.compchemeng.2019.05.029.
URL https://www.sciencedirect.com/science/article/pii/S0098135419300754

[14] J. R. Vázquez-Canteli, Z. Nagy, Reinforcement learning for demand response: A review of algorithms and modeling techniques, Applied Energy 235 (2019) 1072–1089. doi:https://doi.org/10.1016/j.apenergy.2018.11.002.
URL https://www.sciencedirect.com/science/article/pii/S0306261918317082

[15] A. Perera, P. Kamalaruban, Applications of reinforcement learning in energy systems, Renewable and Sustainable Energy Reviews 137 (2021) 110618. doi:https://doi.org/10.1016/j.rser.2020.110618.
URL https://www.sciencedirect.com/science/article/pii/S1364032120309023

[16] D. Cao, W. Hu, J. Zhao, G. Zhang, B. Zhang, Z. Liu, Z. Chen, F. Blaabjerg, Reinforcement learning and its applications in modern power and energy systems: A review, Journal of Modern Power Systems and Clean Energy 8 (6) (2020) 1029–1042. doi:10.35833/MPCE.2020.000552.

[17] Y. Ye, Y. Tang, H. Wang, X.-P. Zhang, G. Strbac, A scalable privacy-preserving multi-agent deep reinforcement learning approach for large-scale peer-to-peer transactive energy trading, IEEE Transactions on Smart Grid 12 (6) (2021) 5185–5200. doi:10.1109/TSG.2021.3103917.

[18] Y. Ye, D. Papadaskalopoulos, Q. Yuan, Y. Tang, G. Strbac, Multi-agent deep reinforcement learning for coordinated energy trading and flexibility services provision in local electricity markets, IEEE Transactions on Smart Grid 14 (2) (2023) 1541–1554. doi:10.1109/TSG.2022.3149266.

[19] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, C. S. Lai, A multi-agent reinforcement learning-based data-driven method for home energy management, IEEE Transactions on Smart Grid 11 (4) (2020) 3201–3211. doi:10.1109/TSG.2020.2971427.

[20] S. Bose, E. Kremers, E. M. Mengelkamp, J. Eberbach, C. Weinhardt, Reinforcement learning in local energy markets, Energy Informatics 4 (1) (2021) 7. doi:10.1186/s42162-021-00141-z.
URL https://doi.org/10.1186/s42162-021-00141-z

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

[22] S. Zhou, Z. Hu, W. Gu, M. Jiang, X.-P. Zhang, Artificial intelligence based smart energy community management: A reinforcement learning approach, CSEE Journal of Power and Energy Systems 5 (1) (2019) 1–10. doi:10.17775/CSEEJPES.2018.00840.

[23] H. Zang, J. Kim, Reinforcement learning based peer-to-peer energy trade management using community energy storage in local energy market, Energies 14 (14) (2021). doi:10.3390/en14144131.
URL https://www.mdpi.com/1996-1073/14/14/4131

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

[25] D. Papadaskalopoulos, G. Strbac, Nonlinear and randomized pricing for distributed management of flexible loads, IEEE Transactions on Smart Grid 7 (2) (2016) 1137–1146. doi:10.1109/TSG.2015.2437795.

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

[27] B. Baker, I. Kanitscheider, T. M. Markov, Y. Wu, G. Powell, B. McGrew, I. Mordatch, Emergent tool use from multi-agent autocurricula, CoRR abs/1909.07528 (2019). arXiv:1909.07528.
URL http://arxiv.org/abs/1909.07528

[28] H.-M. Chung, S. Maharjan, Y. Zhang, F. Eliassen, Distributed deep reinforcement learning for intelligent load scheduling in residential smart grids, IEEE Transactions on Industrial Informatics 17 (4) (2021) 2752–2763. doi:10.1109/TII.2020.3007167.

[29] Q. Zhang, K. Dehghanpour, Z. Wang, F. Qiu, D. Zhao, Multi-agent safe policy learning for power management of networked microgrids, IEEE Transactions on Smart Grid 12 (2) (2021) 1048–1062. doi:10.1109/TSG.2020.3034827.

[30] K. Nweye, S. Sankaranarayanan, Z. Nagy, Merlin: Multi-agent offline and transfer learning for occupant-centric operation of grid-interactive communities, Applied Energy 346 (2023) 121323. doi:https://doi.org/10.1016/j.apenergy.2023.121323.
URL https://www.sciencedirect.com/science/article/pii/S0306261923006876

[31] T. Capper, A. Gorbatcheva, M. A. Mustafa, M. Bahloul, J. M. Schwidtal, R. Chitchyan, M. Andoni, V. Robu, M. Montakhabi, I. J. Scott, C. Francis, T. Mbavarira, J. M. Espana, L. Kiesling, Peer-to-peer, community self-consumption, and transactive energy: A systematic literature review of local energy market models, Renewable and Sustainable Energy Reviews 162 (2022) 112403. doi:https://doi.org/10.1016/j.rser.2022.112403.
URL https://www.sciencedirect.com/science/article/pii/S1364032122003112

[32] N. Liu, X. Yu, C. Wang, C. Li, L. Ma, J. Lei, Energy-sharing model with price-based

demand response for microgrids of peer-to-peer prosumers, IEEE Transactions on Power Systems 32 (5) (2017) 3569–3583. doi:10.1109/TPWRS.2017.2649558.

[33] F. Lezama, J. Soares, P. Hernandez-Leal, M. Kaisers, T. Pinto, Z. Vale, Local energy markets: Paving the path toward fully transactive energy systems, IEEE Transactions on Power Systems 34 (5) (2019) 4081–4088. doi:10.1109/TPWRS.2018.2833959.

[34] R. Ghorani, M. Fotuhi-Firuzabad, M. Moeini-Aghtaie, Optimal bidding strategy of transactive agents in local energy markets, IEEE Transactions on Smart Grid 10 (5) (2019) 5152–5162. doi:10.1109/TSG.2018.2878024.

[35] S. Zhang, D. May, M. Gül, P. Musilek, Reinforcement learning-driven local transactive energy market for distributed energy resources, Energy and AI 8 (2022) 100150. doi:https://doi.org/10.1016/j.egyai.2022.100150.
URL https://www.sciencedirect.com/science/article/pii/S2666546822000118

[36] V. Dudjak, D. Neves, T. Alskaif, S. Khadem, A. Pena-Bello, P. Saggese, B. Bowler, M. Andoni, M. Bertolini, Y. Zhou, B. Lormeteau, M. A. Mustafa, Y. Wang, C. Francis, F. Zobiri, D. Parra, A. Papaemmanouil, Impact of local energy markets integration in power systems layer: A comprehensive review, Applied Energy 301 (2021) 117434. doi:https://doi.org/10.1016/j.apenergy.2021.117434.
URL https://www.sciencedirect.com/science/article/pii/S0306261921008266

[37] S. Zhang, P. Musilek, The impact of battery storage on power flow and economy in an automated transactive energy market, Energies 16 (5) (2023). doi:10.3390/en16052251.
URL https://www.mdpi.com/1996-1073/16/5/2251

[38] D. May, P. Musilek, Transactive local energy markets enable community-level resource coordination using individual rewards, arXiv preprint arXiv:2403.15617 (2024). doi:https://doi.org/10.48550/arXiv.2403.15617.
URL https://arxiv.org/abs/2403.15617

[39] R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction, 2nd Edition, The MIT Press, 2018.
URL http://incompleteideas.net/book/the-book-2nd.html

[40] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, CoRR abs/1707.06347 (2017). arXiv:1707.06347, doi:https:

//doi.org/10.48550/arXiv.1707.06347.
URL http://arxiv.org/abs/1707.06347

[41] J. Schulman, P. Moritz, S. Levine, M. Jordan, P. Abbeel, High-dimensional continuous control using generalized advantage estimation, arXiv preprint arXiv:1506.02438 (2015). doi:https://doi.org/10.48550/arXiv.1506.02438.

[42] F. Christianos, L. Schäfer, S. Albrecht, Shared experience actor-critic for multi-agent reinforcement learning, arXiv preprint arXiv:2006.07169 33 (2020) 10707–10717. doi:https://doi.org/10.48550/arXiv.2006.07169.

[43] J. Schulman, The nuts and bolts of deep rl research, in: NIPS Deep RL Workshop, 2016.
URL http://joschu.net/docs/nuts-and-bolts.pdf

[44] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, N. Dormann, Stable-baselines3: Reliable reinforcement learning implementations, Journal of Machine Learning Research 22 (268) (2021) 1–8.
URL http://jmlr.org/papers/v22/20-1364.html

[45] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780. doi:https://doi.org/10.1162/neco.1997.9.8.1735.
URL https://doi.org/10.1162/neco.1997.9.8.1735

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

[47] M. Andrychowicz, A. Raichuk, P. Stańczyk, M. Orsini, S. Girgin, R. Marinier, L. Hussenot, M. Geist, O. Pietquin, M. Michalski, S. Gelly, O. Bachem, What matters in on-policy reinforcement learning? a large-scale empirical study (2020). arXiv:2006.05990, doi:https://doi.org/10.48550/arXiv.2006.05990.

[48] A. Ilyas, L. Engstrom, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, A. Madry, Are deep policy gradient algorithms truly policy gradient algorithms?, CoRR abs/1811.02553 (2018). arXiv:1811.02553.
URL http://arxiv.org/abs/1811.02553

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

[50] K. Nweye, S. Siva, G. Z. Nagy, The CityLearn Challenge 2022 (2023). doi:10.18738/T8/0YLJ6Q.
URL https://doi.org/10.18738/T8/0YLJ6Q

[51] M. Pleines, M. Pallasch, F. Zimmer, M. Preuss, Generalization, mayhems and limits in recurrent proximal policy optimization (2022). arXiv:2205.11104, doi:https://doi.org/10.48550/arXiv.2205.11104.

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