ORIGINAL RESEARCH **Open access**

A Deep Reinforcement Learning Framework for Transactive Energy Management in Local Energy Markets

Amal Al-Balushi[1*], Saif Al-Hinai[1]

Abstract

The push toward a sustainable energy future, accompanied by the rapid uptake of distributed energy resources, puts growing strain on the reliability and stable operation of the electricity grid. A key facet of this strain is the steep climb in net load fluctuation experienced at the grid edge. Transactive energy, realized through local energy markets, has lately been proposed as a compelling answer, delivering decentralized, indirect demand response at the neighborhood scale. Model-free control strategies, especially deep reinforcement learning (DRL), look well placed to decentralize the automation of market participation. Prior work at the crossroads of transactive energy and model-free control typically highlights socioeconomic gains and self-consumption rates. At the same time, the pressing aim of curbing net load fluctuation across a community remains overlooked. This study remedies that oversight by training a population of deep reinforcement learning agents tasked with automating end-user engagement in an economy-driven, autonomous local energy market (ALEX). In this arrangement, agents operate in complete isolation from one another and care only about shrinking their own electricity bills. The findings expose a strong link between lower bills and smoother net load profiles. The impact on fluctuation is measured over multiple time windows using an open-source dataset and several yardsticks: ramping rate, daily and monthly load factor, and daily average and total peak export and import. To benchmark the proposed DRL scheme, its agents are measured against a near-optimal dynamic programming solution, with a no-intervention baseline for reference. The dynamic programming comparison cuts average daily import, export, and peak demand by 22.05%, 83.92%, and 24.09%, respectively. The RL agents prove equally effective or better, registering reductions of 21.93%, 84.46%, and 27.02% on the same metrics. The upshot is that DRL handles such decentralized grid management tasks with near-optimal performance while scaling naturally.

Keywords Demand response, Distributed energy resources, Reinforcement learning, Deep reinforcement learning, Local energy markets, Distributed energy resource management

*Correspondence:
Amal Al-Balushi
amal.balushi@squ.edu.om
[1] Department of Artificial Intelligence for Sustainable Energy, College of Engineering, Sultan Qaboos University, Muscat, Oman

Introduction

Shifting toward sustainable energy is fundamental in the battle against climate change. In this arena, the meeting of breakneck technological change with sluggish regulatory evolution has triggered a surge in distributed energy resources (DERs), fundamentally altering the landscape at the grid edge where power consumers sit [1]. One consequence is a sharp acceleration in net load fluctuation at the grid edge. The word fluctuation here covers the composite impact of intermittency together with other net load disturbances, such as those stemming from electric vehicle charging. This marked increase makes it ever harder to keep grid operations reliable and efficient [2, 3], pushing a transition toward a Smart Grid that runs autonomously and in a decentralized manner to maintain—and perhaps enhance—operability.

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

To meet these challenges, researchers have poured effort into demand response (DR) techniques. Speaking broadly, DR strategies deploy assorted signals to modulate end-user net load demand in ways that underpin grid efficiency and reliability. Those signals span direct equipment dispatch orders and incentive designs crafted to nudge end-user behavior, drawing a line between direct and indirect DR. The real bottlenecks for indirect DR turn out to be marrying grid-stakeholder objectives with consumer incentives through the right market design, and afterward achieving enough engagement for a meaningful impact [4-6].

Price-schedule-driven methods have long dominated indirect DR practice. Model predictive control (MPC) frameworks have been used to formulate such price schedules dynamically, guided by system state. These approaches build on behavioral models to generate a forecast and then chase optimal load profiles over a future horizon. Still, their foundational need for expert-crafted models, heavy computational footprint, and predisposition toward centralized information handling may undercut their capacity to cope with the fast-moving, heterogeneous transformations now visible at the grid edge.

Stepping in to fill the gaps left by schedule-based thinking, transactive energy (TE) has emerged as a robust alternative. The GridWise Architecture Council [7] defines TE as “the use of a combination of economic and control techniques to improve grid reliability and efficiency”—a definition that resonates strongly with the Smart Grid ethos, where the market serves as a decentralized delivery mechanism for incentives [4].

More recently, the literature has cast Local Energy Markets (LEMs) as a practical vehicle for embedding TE within geographically bounded communities at the grid edge. Mengelkamp *et al*. [8] describe a 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”. LEMs open the door to delivering real-time incentive signals to electricity end-users with the precision and immediacy required within a decentralized framework.

Reviews of completed DR pilot projects confirm what many suspected: automation is not optional if participation rates are to become sufficient [4-6]. Although MPC is well established in the general DR automation literature, model-free contenders such as deep reinforcement learning (DRL) offer a more fitting toolkit for the challenges radiating from the grid edge. Early attention came from high-water-mark DRL victories in games [9-11]; conviction has since deepened with DRL successes in robotics [12] and process control [13], along with a rapidly expanding body of work that brings DRL to bear on electricity grid problems [14-16].

The standout strengths of DRL are model independence, built-in adaptability, scalability, and the ability to keep learning online [13]. In the abstract, these benefits must be balanced against potential drawbacks such as long training times, heavy data appetites, and sample inefficiency. Yet inside a transactive market setting, such worries fade: the market itself ceaselessly churns out data from the ongoing back-and-forth of its participants.

Within this context, a recent wave of studies has set out to automate end-user engagement and DER dispatch within LEMs [17-23], mostly by schooling agents to slash end-user bills through load-shifting capability. Some work has highlighted lower net community energy use [19, 21]; other research has explored unlocking flexibility services [18]. But to the best of the authors’ knowledge, no prior study has yet shown that applying DRL to LEM automation can cut community-level load fluctuation.

Making that case convincingly is no small feat. Even if LEMs are expressly designed to align end-user incentives with grid objectives, incentivized behavior does not automatically tame fluctuation or clean up local power quality [24, 25]. Likewise, the subtle interplay between LEM rules and participant automation may breed unintended side effects [26], a recurring theme when DRL is used to automate complex systems [27].

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Extending the foundations laid by Zhang *et al.* [28] and May and Musilek [29], this article targets the identified knowledge gap by training independent agents to automate end-user market participation and the use of DERs. The chief contributions made here are:

• Shows that community net load fluctuation can be dialed down, revealing that purely selfish bill optimization by agents inside LEMs produces shared benefits that bubble up at the community level.

• Runs the performance evaluation over an open-source dataset, lifting the bar for future benchmarking and making outcomes easier to reproduce and compare.

• Puts DRL agents side by side with several baselines, confirming that the trained agents perform at a level equal to a near-optimal dynamic programming reference.

Related Work and Background

Related literature

The use of DRL in DR, and for electricity grid applications more generally, has attracted intense research interest in recent years [14-16]. Studies that tackle distributed DER coordination via DR outside the LEM framework—such as those by Chung *et al*. [30], Zhang *et al*. [31], and Nweye *et al.* [32]—tend to pursue composite reward formulations and incorporate community-level indicators linked to grid stability or fluctuation, following a direct optimization route.

Meanwhile, the stream of publications centered on LEMs has swelled considerably. Mengelkamp *et al*. [8], Capper *et al.* [33], and Tushar *et al.* [6] provide expansive overviews of the developing LEM landscape. Broadly speaking, this body of work prioritizes the socioeconomic performance of the proposed designs, treating DR dimensions only briefly and keeping performance benchmarking rather circumscribed.

For example, Liu *et al*. [34] advance a LEM-like construct where a supply–demand ratio price coordinates energy movement between microgrids, with MPC driving the automation. Lezama *et al*. [35] likewise inspect LEMs from a grid-integration vantage point, zeroing in on socioeconomic results. Ghorani *et al*. [36] build bidding models for both risk-neutral and risk-averse LEM agents, gauging their socioeconomic value under alternative market designs. Mengelkamp *et al*. [26], for their part, scrutinize different market designs through heuristic agents while keeping the lens fixed on socioeconomic metrics. A growing cluster of studies puts the spotlight on automating LEM participation via DRL. Xu *et al*. [19] apply a MARL Q-learning algorithm to automate engagement in an LEM that distributes a pricing schedule derived from supply and demand forecasts. Zhou *et al.* [22] put forward an economy-driven LEM pricing scheme, refining participant bids through a fusion of Q-learning and fuzzy logic; Zang and Kim [23] similarly coach end-user agents to interface with community-level batteries within LEM environments.

As Mengelkamp *et al.* [26] stress, blending LEMs with automated participation surfaces thorny difficulties and potentially surprising knock-on effects rooted in emergent, intricate system dynamics. Research by Kiedanski *et al*. [24] and Papadaskalopoulos and Strbac [25] indicates that socioeconomic gains in such contexts may not directly carry over into superior grid performance, whether measured by variability dampening or power quality enhancement.

To confront this, some contributions embed electricity grid performance criteria directly into the LEM pricing mechanism or the agent reward signal, departing from LEMs' originally pure economic logic and adopting a direct optimization posture. Chen and Bu [21], for instance, study microgrid trading under LEM conditions with a heavily constrained reward function, showing that the approach lifts self-sufficiency above what expert-designed heuristics and random-action agents achieve in benchmark tests. Ye *et al*. [18] explore the capacity of LEMs to furnish flexibility services, and their work stands apart by measuring against a near-optimal MPC benchmark, establishing a credible performance ceiling. Even here, though, variability-linked metrics that fall outside the agents' explicit optimization targets go unexamined.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

The foundational appeal of LEMs—and of TE more broadly—rests on the conviction that a well-tuned market mechanism can spontaneously encourage a wide range of beneficial behaviors. The deeper hope is to realize this without rigidly lashing the market cost function to those outcomes, thereby enabling lightweight, resilient decentralization that sidesteps the expense of real-time computation of an expressive set of correlated metrics. Put differently, the pure pursuit of lower end-user bills should, in this setting, indirectly and emergently chip away at net load fluctuation. Across the current body of work, however, a demonstration of such emergent behavior using a LEM that relies on DRL for automation has yet to appear. This study is aimed squarely at helping to supply that missing piece.

Autonomous local energy eXchange

The Autonomous Local Energy eXchange (ALEX), originally put forward by Zhang *et al*. [28], functions as a LEM for a community designated as B, within which individual buildings $b \in B$ engage in energy trading steered by a round-based, futures-blind double auction settlement procedure. In a round-based futures market, trading is organized into fixed time intervals. The market takes bids and asks aimed at a future settlement timestep , lodged at the present timestep . Settlement then happens at a later timestep t deliver, where t_{deliver}, with $\begin{matrix} t_{\text{deliver}} \\ > \\ > \end{matrix} \begin{matrix} t_{\text{settle}} \\ t_{\text{now}} \end{matrix}$. Within such a blind double auction, every building interacts with the marketplace blind to the actions of other buildings.

Inside ALEX, participants share no information; each is selfish and chases the smallest attainable electricity bill. A building's electricity bill splits into two main pieces: the market bill and the grid bill. The market bill tallies the total settlement cost by pairing each settlement price with the matching quantity for that building. The grid bill covers the shortfall needed to meet household energy demand, priced at the prevailing grid selling or buying rate according to the applicable net-billing arrangement. A profitability spread between the grid selling and buying prices supplies the pull toward LEM use, rendering any trade routed through the LEM strictly beneficial. Achieving this condition could involve drawing on tracked greenhouse gas emission savings or partial fee relief [33, 37].

Zhang *et al*. [28] dig into the core properties the ALEX settlement mechanism must possess to push RL agents toward learning prices that track the supply–demand ratios prevailing at settlement timestep . Casting ALEX as a competitive mixed-form stochastic game points to the presence of at least one Nash equilibrium.[1] That insight helps uncover a market mechanism carrying the requisite traits, via experiments that deploy tabular Q-learning bandits across varied but static supply–demand ratios. Those experiments then yield a market price function expressed in terms of the current supply–demand ratio. A follow-on paper by Zhang and Musilek [38] examines a system that adds a communal battery energy storage system (BESS) operated by an expert-crafted heuristic, showing that voltage-frequency constraint breaches can be dodged on a test circuit.

May and Musilek [29] go further in probing ALEX as a DR instrument. The authors simulate a group of near-optimal, rational actors on ALEX through an iterative best-response and dynamic programming procedure and then stack their performance against several baselines. The resulting policies lay bare an emergent community-level coordination of DERs, sparked by the incentives woven into the LEM. Tellingly, this coordination comes about even though every participant sees only building-level data and single-mindedly pares down its own electricity bill. Consequently, these policies outstrip the benchmark building-level DR system across an array of community net-load-related metrics that capture net load fluctuation at the community scale. While this agent behavior is encouraging, it relies on a pure search method that presupposes flawless forecasts of end-user generation and demand.

This study aims to push those results further by training a set of DRL agents on the identical task, yet without giving them flawless forecasts, while still reaching a similar level of fluctuation dampening.

Reinforcement learning and Markov decision processes

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Reinforcement Learning (RL) offers a computational framework in which an agent progressively improves how it interacts with an environment by following a reward signal, as illustrated in **Figure 1.**

The underlying mathematical structure is the Markov Decision Process (MDP), described by the tuple (S, A, Pa, Ra). Here, S captures the universe of possible states, A the repertoire of available actions, Pa encodes the odds of moving from state s to a next state s´ after picking action a, and Ra specifies the payoff collected right away. A policy π characterizes an agent's behavioural pattern through a probabilistic link from state s to action a—commonly a Gaussian distribution whose central tendency μ and spread σ are functions of the state s.

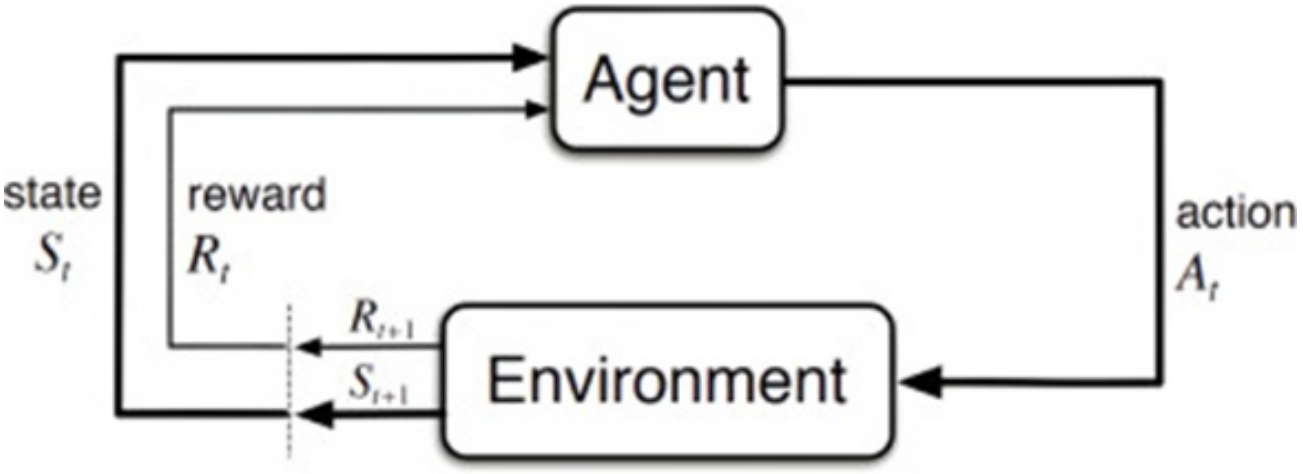

Figure 1. Agent–environment interaction loop, taken from Sutton and Barto [39].

Convention dictates that RL agents operate inside a discrete-time MDP, where successive time steps t trace the arc of interaction from start t = 0 to finish t = T. The return G tallies the sum of discounted future rewards:

$$G_t = \sum_{t=0}^{T} \gamma^t R_{t+1}. \quad (1)$$

Which paves the way for defining state value:

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

And state–action value:

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

With γ playing the role of the discount factor. The overriding goal is to locate an optimal policy π that maximizes the expected return EG.

A key advantage of RL is that optimal policies can be unearthed through firsthand environmental contact. This frees agents to manage unknown transition dynamics and hidden factors without constructing an explicit environment model. In contrast to other MDP search strategies, RL agents gradually distil the optimal policy π through temporal- difference learning and bootstrapping, fine-tuning their parameters θ in response to the rewards they harvest. Updating those parameters relies on the loss function native to the RL algorithm, generally executed through stochastic gradient descent. The algorithmic landscape splits into two traditions: value-based schemes and policy gradient schemes. Value-based

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

approaches concentrate on estimating state values V or state–action values Q, whereas policy gradient approaches tackle policies π directly via a policy loss

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

Actor-critic designs extend policy gradients and keep variance in check by calling upon a critic that estimates state values V and works out advantages A,

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

Deep Reinforcement Learning (DRL), the marriage of RL with deep neural networks, has surged in popularity for cracking complex MDPs [9, 11]. DRL recipes stash agent–environment interactions inside replay buffers to feed multiple rounds of mini-batch gradient updates, making it necessary to tell apart the sampling parameters θold from the freshly updated parameters θ.

Proximal Policy Optimization (PPO), a prominent actor-critic algorithm unveiled by Schulman *et al*. [40], leans on a clipped surrogate objective rooted in the probability ratio r(θ), which measures how far the new policy $\pi_{\theta_{old}}$ has strayed from the old policy π_(θ_old). This brake on policy drift helps keep data collection dependable. The actor loss squeezes the policy ratio r(θ) to stay within a tolerance window ϵ,

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

Most PPO implementations also weave in generalized advantage estimation (GAE) to cut the variance of the advantage A, as pioneered by Schulman *et al*. [41].

Materials and Methods

The ambition of this study is to carry forward the work of Zhang *et al.* [28] and May and Musilek [29] by equipping DRL agents to participate autonomously in ALEX. The anticipated outcome is a level of spontaneous community-level variability reduction that stands shoulder to shoulder with the near-optimal search scheme laid out by May and Musilek [29], but without leaning on an unerring forecast. Putting such variability dampening on display inside a DRL-steered LEM environment would address the conspicuous gap.

To bring this about, the present section reimagines the ALEX environment as a Markov Decision Process (MDP), sets out the DRL algorithm chosen to train the agents and unpacks the experimental blueprint, together with the yardsticks and baselines against which performance is measured.

Autonomous local energy eXchange as a Markov decision process

Reframing ALEX through the lens of an MDP calls for defining what each agent perceives O, what it can do a, what rewards r it chases, and what policy π drives it. The formulation advanced here diverges from the earlier version [29] by embedding modifications tailored to the futures-market character of ALEX. This is essential, given the rule that DRL agents must not draw on information from the future. The design also caters to continuous observation and action spaces for the DRL agents.

The MDP of a single agent encapsulates the vantage point of one participant inside the ALEX ecosystem. Since ALEX participants neither swap information nor communicate, this single-agent MDP remains partially observable, a sharp departure from the fully deterministic joint MDP. In the context of this study, DRL agents are required to function as wholly

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

independent entities, finding their way through a continuous action space paired with a partially observable, continuous observation space. Accordingly, this section adopts the terminology observation space O in place of state space S.

The observation space O^b available to one agent at timestep t gathers several continuous signals: the prevailing net load E_t^b, battery state of charge SoC_t^b, the average price settled most recently $p_{tlast\ settled}^{bid}$, and the aggregate bid and ask volumes from the latest settlement round, $q_{tlast\ settled}^{bid}$ and $q_{tlast\ settled}^{ask}$, respectively. Cyclical time signatures—daily and yearly rhythms—are captured not by the raw timestamp but by sine and cosine transforms of the current timestep t.

$$O_t^b := (\ sin(t)_{year}, cos(t)_{year}, sin(t)_{day}, cos(t)_{day}, E_t^b, SoC_t^b, p_{t_{\text{last settled}}}, q_{t_{\text{last settled}}}^{bid}, q_{t_{\text{last settled}}}^{ask}\ ). \quad (7)$$

Crucially, forward-looking data, most notably the net load expected at settlement time $E_{t_{settle}}^b$, stays outside this observation space. Early-stage testing hints that the feature set outlined here represents the minimal viable configuration for enabling optimal decisions.

Departing from the action space conceived by Zhang *et al.* [28], the action space A^b for an agent at timestep t contains one item only: the continuous battery action, slated for execution at the future settlement timestep t_{settle} and kept within the battery's charge and discharge envelopes. The bid and ask quantities at settlement time t_{settle} follow from the residual net load, while bid and ask market conditions set the prices once the round wraps up, in accordance with the price curve specified by Zhang *et al*. [28].

The building policy π_b emerges from blending two policies:

• The self-sufficiency maximizing, greedy battery policy π0, built exclusively from $\mu 0$.

• The agent's learned policy $\pi\theta$, with mean $\mu\theta$ and standard deviation $\sigma\theta$.

Producing a Gaussian distribution with:

$$\mu_b = \mu_0 + \mu_\theta, \quad (8)$$

$$\sigma_b = \sigma_\theta. \quad (9)$$

This way of structuring actions and agent policy brings a handful of concrete advantages that markedly shorten the learning journey of the DRL agents under investigation. The policy π0 can be assessed at settlement time and furnishes a passable starting heuristic, even though it may be a long way from optimal. The arrangement supports richer state-space exploration while shouldering some of the internal environment modelling burden the agent would otherwise bear. Action constraints are imposed by pushing π_b through a tanh squashing function, in the same spirit as SAC's bounded policy enforcement [42].

The mandate given to agents is to shrink their market-bound electricity bill *bill*^{b}. The reward function is constructed as the negated gap between the electricity bill $bill_{t_{settle}}^b$ and the bill attributable to the self-sufficiency maximizing policy π0, written $bill_{tsettle}^{b,\pi_0}$,

$$r_t^b := bill_{t_{\text{settle}}}^{b,\pi_0} - bill_{t_{\text{settle}}}^b. \quad (10)$$

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

When set beside the plain participant electricity bill as a reward signal, this formulation delivers a far more transparent read on whether the RL agents are genuinely learning something worthwhile: a positive reward value signals that they are pulling ahead of the greedy policy baseline.

Shared experience recurrent proximal policy optimization

Early forays showed that off-policy algorithms along the lines of Soft Actor-Critic (SAC) [42] trailed on-policy alternatives like PPO. The DRL algorithm selected for this investigation is consequently PPO. It is a battle-tested algorithm with a rich ecosystem of well-benchmarked implementations that supply a sturdy platform for extension and adaptation. Stacked against value-based methods, PPO copes better with the continuous action spaces featured in the MDP at hand. Additionally, behaving stochastically can prove advantageous when contending with mixed-form stochastic games such as ALEX. A decisive factor in selecting RL over other plausible methods is the toolkit it provides for grappling with uncertainty in environmental dynamics without building a model.

The agents in this study are schooled as independent agents that pool their experiences [43]. Although each agent operates autonomously and accesses only building-level data, the trajectories they generate are merged in a shared replay buffer. During trajectory collection, the actors run as independent copies of the same actor and critic neural network, which later undergoes updates drawn from the shared replay buffer. This preserves strict independence among agents during rollout while accelerating convergence. Christianos *et al*. [43] chronicled how this design lifts performance in demanding multi-agent domains relative to a regime of wholly independent learning. Observations are standardised and mean-shifted; rewards are standardised only, adhering to the best-practice guidelines of Schulman [44].

The remainder of this section lays out the modifications woven into the base PPO algorithm. A recurrent PPO [45], centred on a Long Short-Term Memory (LSTM) [46] hidden layer that both actor and critic share, is fortified with recurrent burn-in and initialization, as championed by Kapturowski *et al*. [47]. Taking a leaf from Andrychowicz *et al.* [48], after processing a replay buffer, the freshly updated weights θ are deployed to recompute the agent LSTM hidden states across the full trajectory traversed during the current episode. Both tweaks counteract the hazard of hidden-state representations growing stale or drifting, thereby honing the agent's ability to construct meaningful state representations and sustain a long-term context. With Ilyas *et al*. [49] as a guide and convergence toward a Nash equilibrium as the target, the learning rate is progressively dialled down over the training run. Rather than fixing the terminal transition value at T to zero, this study draws it from the critic's own value estimate; preliminary experiments suggest this hastens the critic's convergence to a higher explained variance. Furthermore, rather than bluntly enforcing action-space boundaries by clipping the Gaussian distribution, the algorithm adopted here employs a squashed Gaussian distribution followed by renormalization, a practice made widespread by soft actor-critic algorithms [42].

The weights and biases populating the actor's final layer are sampled uniformly from the interval between 0.001 and −0.001, guaranteeing that the mean μ initially hovers around an expected value of 0. Along parallel lines, the policy's standard deviation σ is configured to start from a very tight, narrow band. Arranged this way, the agent can begin its training journey by drawing on trajectories that cluster near the self-sufficiency maximizing policy π0. Underpinning this tactic is the conjecture that the optimal policy π_(θ)^ is situated far closer to the self-sufficiency policy π0 than it is to a policy driven by pure chance. Drastic deviations from π0 are viewed as highly circumstantial, whereas modest, incremental deviations are treated as the norm. Through the prism of task decomposition, the RL agents tackle two challenges in sequence: first, mastering load shifting to maximise self-sufficiency—a task directed inwards—and subsequently, learning how to engage with the marketplace—a task directed outwards. The design therefore steers the agents toward acquiring load-shifting competence early, deferring market exploitation to a later stage. Both of these choices produced marked convergence uplifts for the problem at hand during preliminary experimentation.

Experimental design

The CityLearn2022 dataset [50] serves as the training ground and testing arena for the DRL agents. Thanks to its open-source nature, future work can perform direct, apples-to-apples benchmarking against the results documented here

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

across a wide variety of DR contexts. Spanning a full year of hourly observations for 17 smart community buildings, the dataset bundles together energy demand profiles, photovoltaic generation traces, and BESS operating characteristics. In keeping with standard conventions adopted across comparable investigations, the broader grid is modelled as an unbounded reservoir, freely absorbing any surplus and supplying any shortfall without constraints at both the prosumer scale and the community scale. A distinct, independently acting agent is assigned to each building and trained according to the procedure. Consequently, one episode corresponds to a single longitudinal pass over the full dataset. It extends for 8760 steps, whereas one run signifies the complete training cycle for the entire cohort of 17 agents. For evaluation purposes, the parameter snapshot θ that attains the peak episodic communal return G^B is extracted from a run, on the premise that this snapshot represents the highest-performing equilibrium struck among the agents. This snapshot is overwritten during training whenever a fresh communal-return record is set. From a total of 5 runs, the run occupying the median position is singled out for benchmarking.

Agent performance is assessed through a battery of metrics borrowed from May and Musilek [29]. Every performance indicator in this study is cast as a function of the community net load E^B, defined straightforwardly as the aggregate across all individual building net loads E^b. In the formulas that follow, n_d denotes the count of days contained in the dataset, d represents the number of time steps packed into a single day, and t marks the present time step. The notations max_{start}^{stop} and min_{start}^{stop} extract the largest and smallest values observed across the span from start to stop, respectively. Because the dataset used here is hourly, the conversion step from kilowatt-hours (kWh) to kilowatts (kW) is left implicit. The full roster of performance metrics includes:

• The average daily imported energy:

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

• The average exported energy:

$$\overline{E}_{d,-} = \frac{-1}{n_d} \sum_{d=0}^{n_d} \left(\sum_{t \in d} \min(E^B(t), 0) \right) \quad (12)$$

• The average daily peak:

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

• The average daily valley:

$$\overline{P}_{d,-} = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(\min_{t \in d} E^B(t) \right) \quad (14)$$

• The absolute maximum peak:

$$P_+ = \max_{t=0}^{T} E^B(t) \quad (15)$$

• The absolute minimum valley:

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

$$P_{-} = \min_{t=0}^{T} E^{B}(t) \quad (16)$$

• The average daily ramping rate:

$$\overline{R}_d = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(\sum_{t \in d} |\nabla E^{B}(t)| \right) \quad (17)$$

• The daily load factor complement:

$$1 - L_d = \frac{1}{n_d} \sum_{d=0}^{n_d} \left(1 - \frac{\text{mean}_{t \in d} E^{B}(t)}{\max_{t \in d} E^{B}(t)} \right) \quad (18)$$

• The monthly load factor complement:

$$1 - L_m = \frac{1}{n_m} \sum_{m=0}^{n_m} \left(1 - \frac{\text{mean}_{t \in m} E^{B}(t)}{\max_{t \in m} E^{B}(t)} \right). \quad (19)$$

Together, these metrics paint a multifaceted picture of how the community net load E^B varies across different time horizons. The spectrum runs from the hour-by-hour scale, encapsulated by the ramping rate $\overline{R}_d$, through the daily and monthly scales, encapsulated by the load factor complements 1 - L_d and 1 - L_m. The yearly and daily averages of peak load demands and generation figures complement this picture by illuminating patterns of community energy consumption and the stress placed on physical infrastructure. A unifying principle across all metrics is that lower numerical readings equate to superior performance. In their totality, these indicators supply a solid evaluative scaffold for judging any DR system's overall influence on net load variability.

To contextualise the achievements of the trained DRL agents, three comparative benchmarks are deployed:

• NoDERMS: This baseline captures the community in its unaltered, default condition, where none of the buildings activate their battery storage capability. It serves as the untouched reference point and is expected to be easily surpassed by any functional DR scheme.

• IndividualDERMS: In this setup, every building within the community follows net-billing logic. Buildings pursue self-sufficiency by trimming building-level peaks and valleys while suppressing the ramping rate [29]. This benchmark provides a credible performance anchor, resembling a carefully tuned heuristic system commonly found in contemporary DR practice. A key distinction from the proposed DRL agents is that this benchmark possesses perfect foresight.

• ALEX DP: This benchmark represents a near-optimal policy operating within a discretised version of ALEX's MDP, uncovered through a dynamic programming search that marries iterative best response with value iteration [29]. A key distinction from the proposed DRL agents is that this benchmark possesses perfect foresight.

The experimental blueprint is deliberately structured to bridge the research gap flagged by scrutinising ALEX RL's results through the lens of the metrics set out above. ALEX RL is weighed against the NoDERMS baseline, with Individual DERMS acting as a sensible performance yardstick and ALEX DP serving as an approximate optimal frontier.

Results and Discussion

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

This study is driven by the ambition to close a conspicuous gap in the literature—identified in the background discussion—by showing that DRL agents embedded within a LEM can drive down community-level net load variability. In pursuit of that goal, this section first forges a clear empirical link between participant bill reductions and performance metrics inside ALEX, working under the assumption of fully converged actors. A head-to-head appraisal of the DRL agents against the benchmarks outlined earlier then follows.

The training philosophy centres squarely on the agents' incremental improvement over the self-sufficiency maximizing policy π0. Convergence trajectories are rendered visible in **Figure 2**, which tracks the average building bill savings logged by ALEX RL across successive episodes, juxtaposed against those of ALEX DP. The shaded envelope captures the dispersion observed across runs.

As **Figure 2** clearly reveals, ALEX RL achieves bill savings that marginally outstrip those delivered by ALEX DP. A subtle but important distinction warrants emphasis: ALEX DP restricts its search to a one-day-ahead window, whereas ALEX RL operates under no such limitation on the temporal reach of its load shifting. These outcomes hint that the CityLearn 2022 community harbours substantial scope for redistributing load across multi-day intervals.

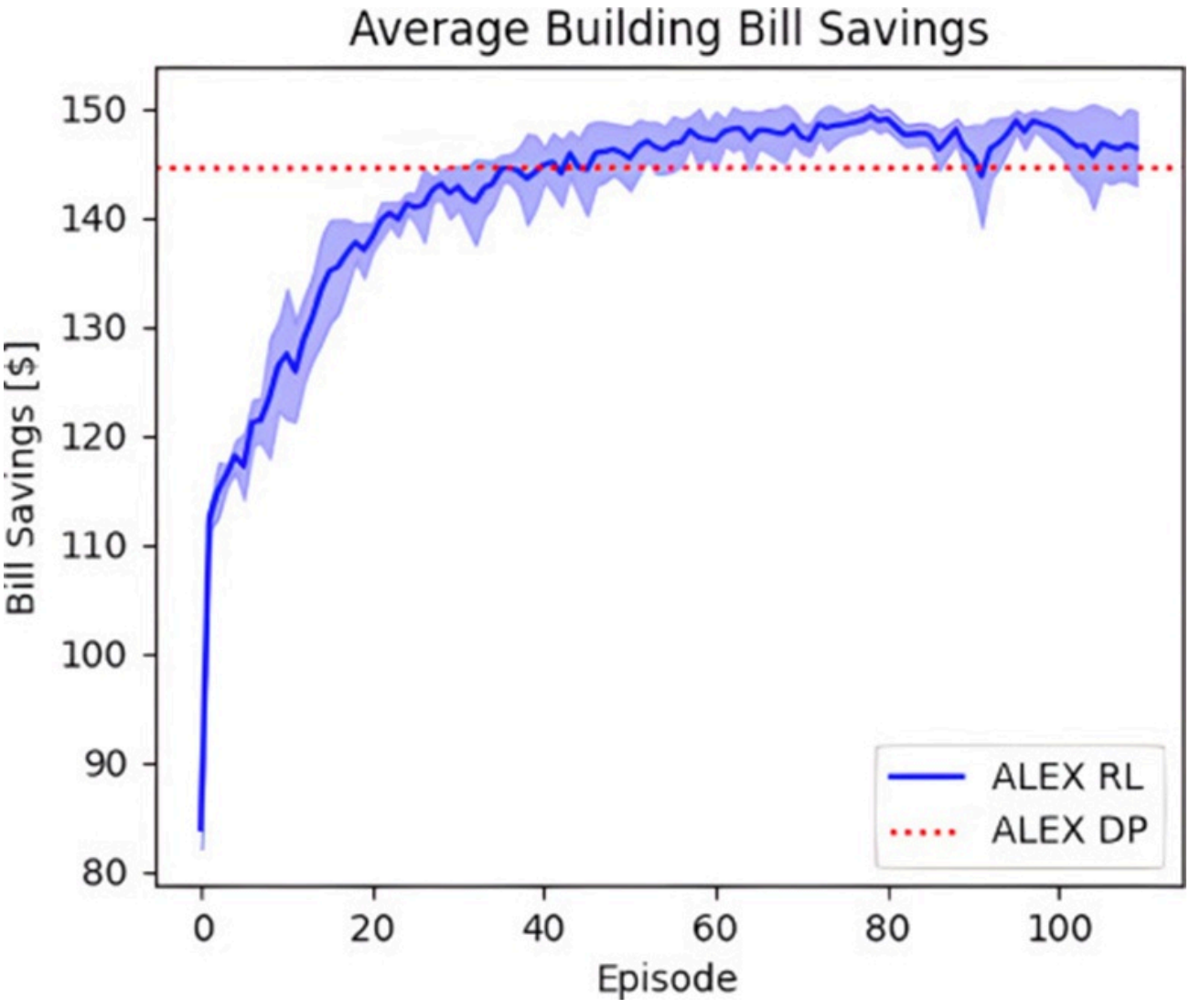

Figure 2. Average participant bill savings comparison between ALEX RL (blue) and ALEX DP (red). Shaded areas depict variance bands between a set of 5 ALEX RL runs, trained over 117 episodes (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

To more robustly tie the achieved bill savings to the evaluation metrics, **Figure 3** plots how the median-performing ALEX RL run evolves on each performance indicator as training progresses. A pronounced downward trajectory is visible for every metric, underscoring a strong link between self-centred bill minimisation and the chosen set of performance measures.

Quantitative scrutiny, collated in **Table 1**, uniformly corroborates that the performance metrics and participant bill savings move together. These observations validate that grooming DRL agents within ALEX nurtures behaviour that, as a by-product, reins in community net load fluctuations.

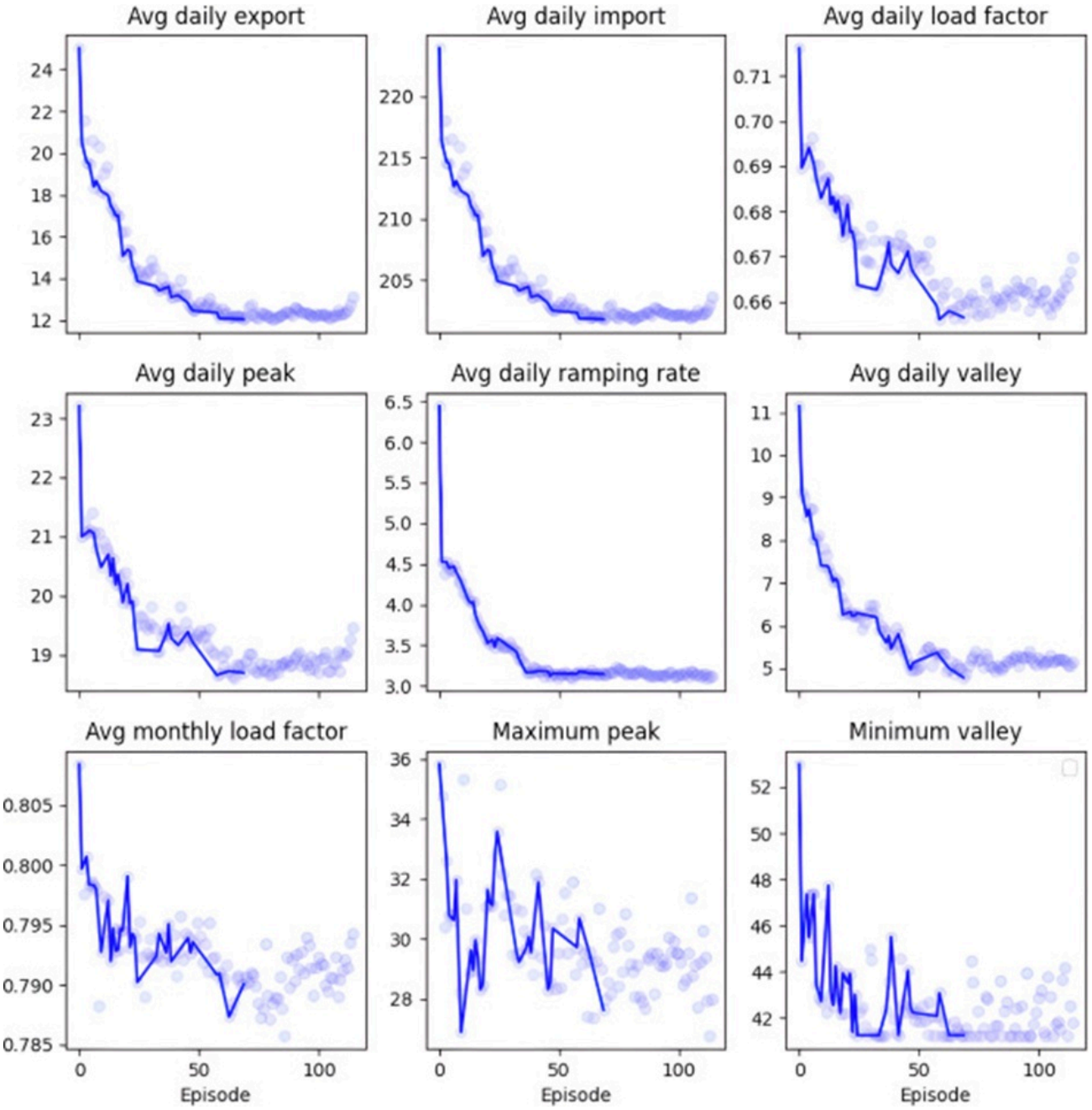

Figure 3. Community-level metric values logged per episode over the training timeline. The faint scattered markers denote individual episode snapshots. At the same time, the solid blue curve traces the metric score linked to the best

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

return registered up to that episode (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

The data suggest that the association linking metrics to return holds steady from run to run. Furthermore, the correlations drawn at the point of peak return consistently land above their episode-by-episode analogues. This pattern is not especially puzzling when one considers that ALEX functions as a mixed-form stochastic game; it plausibly reflects the path taken while settling toward a Nash equilibrium. Put simply, episodes that harvest richer returns tend to coincide with agent policies that sit closer to a mutual best-response posture.

Table 1. Pearson correlation coefficients connecting metrics with realised bill savings; the rightmost column covers Maximum Return episodes paired with their simultaneous metric scores, the middle column covers episodic returns paired with their concurrent episodic metric scores; numbers in brackets give the mean correlation averaged across 5 training runs, while the unbracketed numbers give the correlation for the run that yields the median return.

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

Table 2 sets the median-performing DRL cohort side by side with the nominated benchmarks. Driven by a substantially more intense use of locally sourced energy, average daily import ($\overline{E}_{d,+}$) and export ($\overline{E}_{d,-}$) both contract by 21.9% and 84.4% respectively. At the same time, self-organised peak-shaving behaviour depresses the average daily peak ($\overline{P}_{d,+}$) and valley ($\overline{P}_{d,-}$) by 27.0% and 71.1%, with the absolute maximum peak (P+) and absolute minimum valley (P-) shrinking by 16.0% and 27.0% as well. This behaviour also irons out abrupt swings in community net-load, delivering a 26% cut in the ramping rate ($\overline{R}_d$) and a softening of the broader community net-load envelope that reduces the daily load factor complement (**1 - L_d**) and monthly load factor complement (**1 - L_m**) by 11.0% and 3.6%, in that order. In a nutshell, ALEX RL markedly dampens the footprint of community-level variability across the entire metric suite.

A closer look across benchmarks reveals that the DRL agents comprehensively eclipse IndividualDERMS and notch partial wins over ALEX DP. The ramping rate ($\overline{R}_d$) registers as one area where ALEX RL lags behind both IndividualDERMS and ALEX DP. Another point worth spotlighting is the average daily valley ($\overline{P}_{d,-}$), where ALEX RL posts a value notably higher than ALEX DP—an outcome that catches some attention. IndividualDERMS and ALEX DP both plot their course against a flawless forecast. In contrast, ALEX RL operates without any window into the future and

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

must internally piece together a working approximation of participant net load dynamics. As the metric most directly tuned to near-term swings, the ramping rate ($\overline{R}_d$) picks up the greatest sensitivity to any such misestimates. The divergence in average daily valley ($\overline{P}_{d,-}$) between ALEX RL and ALEX DP may owe something to a calculated strategic tilt. When armed only with an imperfect internal model, biasing toward selling into the grid rather than drawing from it could be the economically less risky play for the DRL agents. Such a scenario might materialise when the market sees bid quantities overwhelming ask quantities, with leftover residual load then settled through standard net-billing arrangements.

Table 2. Synopsis of metrics over a full simulation pass on the CityLearn2022 dataset [50] for NoDERMS, IndividualDERMS, ALEX DP, and ALEX DRL configurations. Values for NoDERMS, IndividualDERMS, and ALEX DP are sourced from May and Musilek [29]. Top scores appear in bold.

Performance metric	Symbol	ALEX RL	ALEX DP	IndividualDERMS	NoDERMS
Mean Daily Energy Import (kWh)	$\overline{E}_{d,+}$	201.83	202.68	214.81	258.54
Mean Daily Energy Export (kWh)	$\overline{E}_{d,-}$	−12.04	−12.46	−26.49	−77.48
Average Daily Peak Demand (kW)	$\overline{P}_{d,+}$	18.69	19.44	19.95	25.61
Average Daily Minimum Demand (kW)	$\overline{P}_{d,-}$	−4.78	−1.67	−6.35	−16.55
Highest Recorded Peak Power (kW)	P_+	41.22	42.37	42.37	49.06
Lowest Recorded Valley Power (kW)	P_-	−27.62	−29.34	−36.80	−37.86
Mean Daily Ramping Rate (kW)	$\overline{R}_d$	3.15	2.84	2.87	4.28
Average Daily Load Factor	$1 - L_d$	0.65	0.64	0.65	0.73
Average Monthly Load Factor	$1 - L_m$	0.79	0.78	0.80	0.82

What also emerges is that ALEX RL seemingly compensates for the lack of an ideal internal model by exploiting its greater latitude to shift load across longer time spans than ALEX DP, squeezing the maximum peak (P+) and minimum valley (P-) down a notch further. Consequently, ALEX RL's superiority on the bill-savings front does not map directly onto across-the-board dominance over ALEX DP on the evaluation metrics. Taken as a whole, ALEX RL's track record sits very close to that of ALEX DP, signalling a broadly equivalent degree of spontaneous, community-level DER orchestration. The weight of the accumulated evidence builds a persuasive narrative that automated participation steered by DRL agents inside a LEM can yield emergent, community-level variability mitigation, effectively sealing the gap that the literature had left open.

In closing, the results drive home the value of fielding DRL agents inside LEMs for load optimisation and point toward significant headroom for taming variability and streamlining energy use at the community scale. The study presented here keeps a deliberately narrow focus, centring on benchmarking a collection of DRL agents within ALEX against a specific dataset. The framework it lays out supplies a decentralised, data-driven scaffold that holds promise for extension to larger, more heterogeneous energy communities. Delving into the complexities that such scaling would introduce, however, sits outside the boundary of the current work.

Conclusion

This study examines how engagement in economy-driven LEMs can be automated through DRL agents.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

The breakneck spread of DERs at the grid edge is amplifying variability and dispersion in community net load, placing fresh strains on electricity grid operations. TE-based DR, enacted through community LEMs, is carving out a role as a practical remedy by dovetailing the interests of electricity end-users with those of grid operators [8, 33, 37]. At the same time, the record of DR pilot programmes drives home that automation is not optional [4, 5]. In step with the decentralised, scattered nature of this challenge, model-free control strategies—with DRL at the forefront—have surfaced as strong contenders for automating LEM participation [17-23]. Although earlier work in this vein has tended to fixate on socioeconomic yardsticks and community net load consumption, a crisp demonstration of variability or dispersion reduction has until now remained elusive.

This paper closes that gap by putting to work a shared-experience [43], recurrent PPO [45] algorithm enriched with several customisations [47-49] to school a set of DRL agents inside ALEX, an economy-driven LEM where each participant selfishly pursues bill minimisation in the absence of information sharing [28]. The polished DRL agents are stacked against benchmarks that include a building-level DR scheme and a near-optimal solution rooted in dynamic programming [29]. Evaluation draws on a collection of metrics that span net load dispersion over multiple time horizons—ramping rate, daily and monthly load factor, peak and average daily import and export among them. The experiments lay bare a strong alignment between relative bill savings and gains on the scrutinised metrics. The trained DRL agents deliver a solid performance, drawing close to—and occasionally besting—the yardsticks set by the near-optimal dynamic programming approach, all while comfortably outpacing the building-level DR strategy.

Directions for follow-up research should centre on probing the portability, generalisability, and scalability of the framework introduced here. One strand involves firming up a sharper performance upper bound by crafting more advanced DRL algorithms expressly tuned to the mixed-form stochastic game profile of LEMs such as ALEX. While this contribution leans on RL agents taught to pursue reward maximisation, recasting the problem as a Constrained MDP [51] could serve to uphold constraint boundaries. Other promising paths include safety-conscious RL [52] and controller-mediated intervention (as in Learning from Intervention [53]) to catch and block hazardous agent choices. Investigating scalability should also confront challenges like escalating computational expense and the demands of coordinating larger agent collectives, alongside the interactions that might unfold among multiple ALEX instances under assorted grid conditions. By the same token, there is scope to explore ALEX's capacity to assimilate further DER types—electric vehicles and renewables such as wind—potentially unlocking broader environmental dividends including emission reductions. Any eventual transition of the technology platform into practice would also need to reckon with the heterogeneous regulatory terrain overseen by grid stakeholders and the attendant social implications.

Beyond this, extending the benchmarking exercise undertaken here to a richer palette of LEM designs and agent heuristics could sharpen understanding of what factors shape the success—or failure—of incentive structures in coaxing out desirable behaviours within these systems [24, 25].

Acknowledgements

This research has been supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada grant RGPIN-2024-04565 and by the NSERC/Alberta Innovates grant ALLRP 561116-20. Part of this work has taken place in the Intelligent Robot Learning (IRL) Lab at the University of Alberta, which is supported in part by research grants from the Alberta Machine Intelligence Institute (Amii), Canada; a Canada CIFAR AI Chair, Amii; Digital Research Alliance of Canada; Huawei; Mitacs, Canada; and NSERC, Canada.

Conflict of interest

None

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Financial support

None

Ethics statement

None

Received: 30 May 2022 Revised: 04 Aug 2022 Accepted: 19 Aug 2022
Published online: 15 November 2023

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. International Renewable Energy Agency. Global energy transformation: A roadmap to 2050 [Technical Report]. 2019th ed. International Renewable Energy Agency; 2019. Available from: https://www.irena.org/publications/2019/Apr/Global-energy-transformation-A-roadmap-to-2050-2019Edition

2. Kok K, Widergren S. A society of devices: Integrating intelligent distributed resources with transactive energy. IEEE Power Energy Mag. 2016;14(3):34-45. https://doi.org/10.1109/MPE.2016.2524962.

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

4. Chen S, Liu CC. From demand response to transactive energy: state of the art. J Mod Power Syst Clean Energy. 2017;5(1):10-9. https://doi.org/10.1007/s40565-016-0256-x.

5. Federal Energy Regulatory Commission. Assessment of demand response and advanced metering [Technical Report]. Federal Energy Regulatory Commission; 2015. Available from: https://www.ourenergypolicy.org/wp-content/uploads/2015/12/demand-response.pdf

6. Tushar W, Yuen C, Saha TK, Morstyn T, Chapman AC, Alam MJE, et al. Peer-to-peer energy systems for connected communities: A review of recent advances and emerging challenges. Appl Energy. 2021;282:116131. https://doi.org/10.1016/j.apenergy.2020.116131.

7. Melton RB. Gridwise transactive energy framework [Technical Report]. The GridWise Architecture Council; 2019.

8. Mengelkamp E, Diesing J, Weinhardt C. Tracing local energy markets: A literature review. it - Inf Technol. 2019;61(2-3):101-10. https://doi.org/10.1515/itit-2019-0016.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

9. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature. 2015;518(7540):529-33. https://doi.org/10.1038/nature14236.

10. Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science. 2018;362(6419):1140-4. https://doi.org/10.1126/science.aar6404.

11. OpenAI, Berner C, Brockman G, Chan B, Cheung V, Debiak P, et al. Dota2 with large scale deep reinforcement learning. arXiv preprint arXiv:1912.06680. 2019. https://doi.org/10.48550/arXiv.1912.06680.

12. Singh B, Kumar R, Singh VP. Reinforcement learning in robotic applications: a comprehensive survey. Artif Intell Rev. 2022:1-46. https://doi.org/10.1007/s10462-021-09997-9.

13. Shin J, Badgwell TA, Liu KH, Lee JH. Reinforcement learning – overview of recent progress and implications for process control. Comput Chem Eng. 2019;127:282-94. https://doi.org/10.1016/j.compchemeng.2019.05.029.

14. Vázquez-Canteli JR, Nagy Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques. Appl Energy. 2019;235:1072-89. https://doi.org/10.1016/j.apenergy.2018.11.002.

15. Perera A, Kamalaruban P. Applications of reinforcement learning in energy systems. Renew Sustain Energy Rev. 2021;137:110618. https://doi.org/10.1016/j.rser.2020.110618.

16. Cao D, Hu W, Zhao J, Zhang G, Zhang B, Liu Z, et al. Reinforcement learning and its applications in modern power and energy systems: A review. J Mod Power Syst Clean Energy. 2020;8(6):1029-42. https://doi.org/10.35833/MPCE.2020.000552.

17. Ye Y, Tang Y, Wang H, Zhang XP, Strbac G. A scalable privacy-preserving multi-agent deep reinforcement learning approach for large-scale peer-to-peer transactive energy trading. IEEE Trans Smart Grid. 2021;12(6):5185-200. https://doi.org/10.1109/TSG.2021.3103917.

18. Ye Y, Papadaskalopoulos D, Yuan Q, Tang Y, Strbac G. Multi-agent deep reinforcement learning for coordinated energy trading and flexibility services provision in local electricity markets. IEEE Trans Smart Grid. 2023;14(2):1541-54. https://doi.org/10.1109/TSG.2022.3149266.

19. Xu X, Jia Y, Xu Y, Xu Z, Chai S, Lai CS. A multi-agent reinforcement learning-based data-driven method for home energy management. IEEE Trans Smart Grid. 2020;11(4):3201-11. https://doi.org/10.1109/TSG.2020.2971427.

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

21. Chen T, Bu S. Realistic peer-to-peer energy trading model for microgrids using deep reinforcement learning. In: 2019 IEEE PES Innovative Smart Grid Technologies Europe. 2019. p. 1-5. https://doi.org/10.1109/ISGTEurope.2019.8905731.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

22. Zhou S, Hu Z, Gu W, Jiang M, Zhang XP. Artificial intelligence based smart energy community management: A reinforcement learning approach. CSEE J Power Energy Syst. 2019;5(1):1-10. https://doi.org/10.17775/CSEEJPES.2018.00840.

23. Zang H, Kim J. Reinforcement learning based peer-to-peer energy trade management using community energy storage in local energy market. Energies. 2021;14(14). https://doi.org/10.3390/en14144131.

24. Kiedanski D, Kofman D, Maillé P, Horta J. 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. 2020. p. 1-7. https://doi.org/10.1109/SmartGridComm47815.2020.9302939.

25. Papadaskalopoulos D, Strbac G. Nonlinear and randomized pricing for distributed management of flexible loads. IEEE Trans Smart Grid. 2016;7(2):1137-46. https://doi.org/10.1109/TSG.2015.2437795.

26. Mengelkamp E, Staudt P, Garttner J, Weinhardt C. Trading on local energy markets: A comparison of market designs and bidding strategies. In: 2017 14th International Conference on the European Energy Market. 2017. p. 1-6. https://doi.org/10.1109/EEM.2017.7981938.

27. Baker B, Kanitscheider I, Markov TM, Wu Y, Powell G, McGrew B, et al. Emergent tool use from multi-agent autocurricula. CoRR abs/1909.07528. 2019. Available from: http://arxiv.org/abs/1909.07528

28. Zhang S, May D, Gül M, Musilek P. Reinforcement learning-driven local transactive energy market for distributed energy resources. Energy AI. 2022;8:100150. https://doi.org/10.1016/j.egyai.2022.100150.

29. May D, Musilek P. Transactive local energy markets enable community-level resource coordination using individual rewards. arXiv preprint arXiv:2403.15617. 2024. https://doi.org/10.48550/arXiv.2403.15617.

30. Chung HM, Maharjan S, Zhang Y, Eliassen F. Distributed deep reinforcement learning for intelligent load scheduling in residential smart grids. IEEE Trans Ind Inf. 2021;17(4):2752-63. https://doi.org/10.1109/TII.2020.3007167.

31. Zhang Q, Dehghanpour K, Wang Z, Qiu F, Zhao D. Multi-agent safe policy learning for power management of networked microgrids. IEEE Trans Smart Grid. 2021;12(2):1048-62. https://doi.org/10.1109/TSG.2020.3034827.

32. Nweye K, Sankaranarayanan S, Nagy Z. Merlin: Multi-agent offline and transfer learning for occupant-centric operation of grid-interactive communities. Appl Energy. 2023;346:121323. https://doi.org/10.1016/j.apenergy.2023.121323.

33. Capper T, Gorbatcheva A, Mustafa MA, Bahloul M, Schwidtal JM, Chitchyan R, et al. Peer-to-peer, community self-consumption, and transactive energy: A systematic literature review of local energy market models. Renew Sustain Energy Rev. 2022;162:112403. https://doi.org/10.1016/j.rser.2022.112403.

34. Liu N, Yu X, Wang C, Li C, Ma L, Lei J. Energy-sharing model with price-based demand response for microgrids of peer-to-peer prosumers. IEEE Trans Power Syst. 2017;32(5):3569-83. https://doi.org/10.1109/TPWRS.2017.2649558.

35. Lezama F, Soares J, Hernandez-Leal P, Kaisers M, Pinto T, Vale Z. Local energy markets: Paving the path toward fully transactive energy systems. IEEE Trans Power Syst. 2019;34(5):4081-8.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

https://doi.org/10.1109/TPWRS.2018.2833959.

36. Ghorani R, Fotuhi-Firuzabad M, Moeini-Aghtaie M. Optimal bidding strategy of transactive agents in local energy markets. IEEE Trans Smart Grid. 2019;10(5):5152-62. https://doi.org/10.1109/TSG.2018.2878024.

37. Dudjak V, Neves D, Alskaif T, Khadem S, Pena-Bello A, Saggese P, et al. Impact of local energy markets integration in power systems layer: A comprehensive review. Appl Energy. 2021;301:117434. https://doi.org/10.1016/j.apenergy.2021.117434.

38. Zhang S, Musilek P. The impact of battery storage on power flow and economy in an automated transactive energy market. Energies. 2023;16(5). https://doi.org/10.3390/en16052251.

39. Sutton RS, Barto AG. Reinforcement learning: an introduction. 2nd ed. The MIT Press; 2018. Available from: http://incompleteideas.net/book/the-book-2nd.html

40. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. CoRR abs/1707.06347. 2017. https://doi.org/10.48550/arXiv.1707.06347.

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

42. Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Dy J, Krause A, editors. Proceedings of the 35th International Conference on Machine Learning. Vol. 80. PMLR; 2018. p. 1861-70. Available from: https://proceedings.mlr.press/v80/haarnoja18b.html

43. Christianos F, Schäfer L, Albrecht S. Shared experience actor-critic for multi-agent reinforcement learning. arXiv preprint arXiv:2006.07169. 2020. p. 10707-17. https://doi.org/10.48550/arXiv.2006.07169.

44. Schulman J. The nuts and bolts of deep RL research. In: NIPS Deep RL Workshop. 2016. Available from: http://joschu.net/docs/nuts-and-bolts.pdf

45. Raffin A, Hill A, Gleave A, Kanervisto A, Ernestus M, Dormann N. Stable-baselines3: Reliable reinforcement learning implementations. J Mach Learn Res. 2021;22(268):1-8. Available from: http://jmlr.org/papers/v22/20-1364.html

46. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735-80. https://doi.org/10.1162/neco.1997.9.8.1735.

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

48. Andrychowicz M, Raichuk A, Stańczyk P, Orsini M, Girgin S, Marinier R, et al. What matters in on-policy reinforcement learning? A large-scale empirical study. arXiv:2006.05990. 2020. https://doi.org/10.48550/arXiv.2006.05990.

49. Ilyas A, Engstrom L, Santurkar S, Tsipras D, Janoos F, Rudolph L, et al. Are deep policy gradient algorithms truly policy gradient algorithms? CoRR abs/1811.02553. 2018. Available from: http://arxiv.org/abs/1811.02553

50. Nweye K, Siva S, Nagy GZ. The CityLearn challenge 2022. 2023.

51. Altman E. Constrained Markov decision processes. Chapman and Hall; 1999.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

52. Hans A, Schneegass D, Schäfer A, Udluft S. Safe exploration for reinforcement learning. 2008. p. 143-8.

53. Saunders W, Sastry G, Stuhlmüller A, Evans O. Trial without error: Towards safe reinforcement learning via human intervention. CoRR abs/1707.05173. 2017. Available from: http://arxiv.org/abs/1707.05173

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press