Consumer-Centric Home Energy Management System Using Trust Region Policy Optimization-Based Multi-Agent Deep Reinforcement Learning

Kuthsav Thattai*, Jayashri Ravishankar, and Chaojie Li
School of Electrical Engineering and Telecommunications
University of New South Wales
Sydney, Australia
ORCID*: https://orcid.org/0000-0002-6475-9177

***Abstract*—Autonomous home energy management system (HEMS) is the key to improving energy efficiency in the active distribution network. This HEMS also needs to maintain customer satisfaction while maximizing cost savings under dynamic price conditions, incorporating uncertainties of consumer behavior, and renewable energy generation. In this paper, a consumer-centric HEMS using Trust Region Policy Optimization (TRPO) based multi-agent deep reinforcement learning (DRL) is presented. This Multi-Agent TRPO (MA-TRPO) based HEMS is trained to respond to the dynamic retail price and the local energy generation by scheduling the Interruptible-Deferrable load (IDA) and Battery Energy Storage System (BESS). Five-minute retail electricity price derived from wholesale market price and the PV generation data derived from real-world PV profiles are used to train the proposed MA-TRPO-based HEMS in discrete action space. The performance of the proposed HEMS is relatively better than the existing policy-gradient-based on-policy approaches such as Proximal Policy Optimization and Policy Gradient-based HEMS as validated via training and testing using the same dataset.**

Keywords— Deep reinforcement learning, Energy Management System, Multi-Agent, Smart Home, Trust region policy optimization

I. Introduction

Gradual decommissioning of fossil-fuel-based conventional power generation and increasing penetration of renewables are making the modern electricity grid more complex and less reliable due to lack of inertia. The key challenges of the increasing proliferation of intermittent renewable energy sources in low voltage distribution networks (LVDNs) include voltage rise, reverse power flow, voltage fluctuations, and unbalance [1]. In the year 2021, nearly 38% of the energy consumption in the US is attributed to residential consumption [2]. Active control options are thus increasingly necessary at all levels of the electricity network within the modern electricity grid. Demand response is one way to exercise such control. Home energy management systems (HEMS) enable active consumer participation via demand response programs [3] and play an important role in enabling control at the LVDN level using economic signals such as 5-minute pricing.

With advances in information and communication technology, especially the evolution of the Internet of Things, wireless sensor networks, etc., remote and/or automated control of household appliances and energy resources have become easier than ever before. Within the framework of economic signals and user comfort, smart and automated HEMS can coordinate and manage local resources. The main objective of the HEMS is to optimize appliances used to minimize energy costs and maintain consumer comfort while responding to the electricity price [4]. To achieve this objective, HEMS monitors and controls the home appliances. Home appliances can be classified into Controllable, Shiftable, and Non-controllable base loads. Further Shiftable/Deferrable home loads can be categorized into non-interruptible deferrable load and interruptible-deferrable load (IDL) [5]. The scope of this research is limited to non-controllable base loads and Interruptible-deferrable loads. This is based on the hypothesis that for an operation profile of an IDA appliance with a high consumer utility preference factor, the HEMS agent learns to minimize the interruptions to avoid penalties associated with delayed start and completion of the task. Various HEMS strategies are discussed in the literature, and a detailed review of recent techniques adopted for HEMS is presented in [6].

Traditional techniques for HEMS presented in the literature include mixed integer linear programming (MILP) [7], mixed integer non-linear programming (MINLP) [8], dynamic programming (DP) [9], stochastic programming (SP) [10], and robust programming (RP) [11]. These traditional techniques are highly dependent on the model and scenario and hence may be unrealistic for highly dynamic and uncertain home environments [12].

Reinforcement learning (RL) falls under the heuristic and meta-heuristic categories of the HEMS techniques [6]. RL is a model-free machine learning approach that does not require prior domain knowledge and can be a powerful tool to schedule energy systems while dealing with uncertainty in local energy generation, and consumption [12].

Deep reinforcement learning (DRL) algorithms can be broadly classified as on-policy and off-policy. On-policy algorithms evaluate and improve the current policy that is taking the action, whereas in off-policy DRL the policy being evaluated and improved is different from the current policy used to take action [13].

Deep Q-Network (DQN) [14], Double Deep Q-Network (DDQN) [15], Deep-Deterministic policy gradient (DDPG) [16], Twin Delayed Deterministic Policy Gradient (TD3) [17], etc., are some of the off-policy DRL algorithms commonly studied for HEMS. Some of the on-policy DRL algorithms used in HEMS include Advantage Actor-Critic (A2C) [18], Proximal Policy Optimization (PPO) [19], and Trust Region Policy Optimization (TRPO) [20]. A recent literature survey in [21] shows that the popularity of PPO was more than TRPO and this is attributed to the recency of TRPO. The initial proposal of TRPO, as outlined in [22], aimed to address decline in performance by solving an optimization problem. This involved placing constraints on the Kullback-Leibler divergence (KL-divergence) between the prior and present

2023 IEEE Belgrade PowerTech | 978-1-6654-8778-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/POWERTECH55446.2023.10202803

policy. This approach ensured that the new policy remained within the trust region while minimizing any drops in the performance. Further, to reduce the variance while tolerable bias is maintained in TRPO, a generalized advantage estimator (GAE) was proposed in [23].

A HEMS using the former TRPO approach is proposed in [20] where a single TRPO agent is modeled and trained to handle both discrete and continuous actions to jointly optimize the scheduling of various types of appliances under uncertainty of price and outdoor temperature. The effectiveness of TRPO-based HEMS was validated and was proved to perform better than other DRLs such as DQN and DDPG. To the best knowledge of the authors, this is the only existing literature where TRPO applications for HEMS are studied. However, local generation sources such as PV are not considered here. Further, the effectiveness of TRPO in multi-agent HEMS is not studied.

In this article, we propose and study a multi-agent GAE-based TRPO for scheduling household appliances in the presence of uncertainties associated with price, local generation, and base load. The proposed HEMS exercises discrete control only and is benchmarked against on-policy DRL algorithms like Policy Gradient (also referred to as REINFORCE) with baseline and PPO. Section II of the paper provides an overview of the smart home environment followed by a discussion on the Markov-Decision Process (MDP) formulation and DRL algorithm in sections III and IV. In section V experimental setup, datasets, and benchmarks are discussed. Results are presented in section VI and the conclusion of this article is presented in section VII.

II. System Overview

The smart home considered in this paper includes a non-controllable base load, interruptible deferrable appliance, and battery energy storage system (BESS). Models of different elements considered in this paper are presented in the following subsections. The simple home environment considered in this paper is presented in Fig. 1.

A. Non-controllable base load (BL)

Base load in the context of this paper is the aggregate of all the loads that are not controllable, non-deferrable, and non-interruptible. If the power consumed by the base load at any given instance is taken as $P_{Bl}(t)$ then base load can be represented by its energy consumption over five minutes since the last action ($t\epsilon[t'-300,t']$).

$$E_{Bl,t} \ = \ \int_{t'-300}^{t'} P_{Bl}(t)\,dt \tag{1}$$

B. Interruptible-Deferrable Appliance

Interruptible-deferrable appliance (IDA) can be time-shifted and interrupted within the preferred operational time $[Sch_{On}, Sch_{Off}]$. During preferred operation time IDA can either be "On" or "Off". The state of the IDA during the preferred operation time can be represented by a control variable $\alpha'_{IDL,t}$ and each control period is for 300 s. For each i^{th} operation profile of IDA, the consumer can assign a preference factor $\varphi_{IDA,\,i}$ such that $0 \le \varphi_{IDA,\,i} \le 1$. A preference factor of 0 ($\varphi_{IDA,\,i}= 0$) means the customer prefers cost savings and 1 ($\varphi_{IDA,\,i}= 1$) means the customer prefers the utility/operation of the appliance. The ID appliance can now be represented using the following expressions:

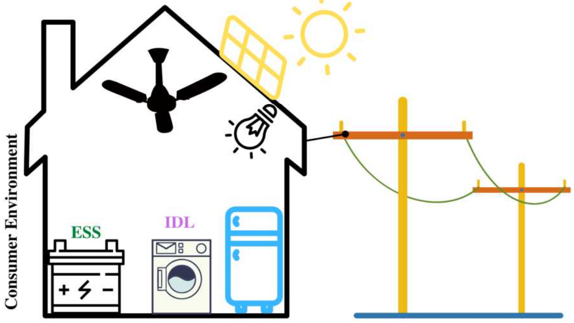

Fig. 1. Simple smart home environment

$$E_{IDA,\,i,\,t} = \begin{cases} \int_{t'-300}^{t'} P_{IDA,\,i}(t)\,dt\,,\ if\ \alpha'_{IDA,\,t}{=}1\ \forall t\epsilon\left[Sch_{On}, Sch_{Off}\right] \\ \qquad\qquad 0,\ Otherwise \end{cases} \tag{2}$$

C. Battery Energy Storage System (BESS)

Battery Energy Storage systems can act as both dispatchable load and source. The scope of this article is limited by considering a fixed voltage model of BESS. BESS can either supply energy by discharging, staying idle, or consuming energy during charging. The state of the battery can be defined using the control variable $\alpha'_{BESS,\,t}$.

$$\alpha'_{BESS,\,t} = \begin{cases} 1,\ Discharge \\ 0,\ \ Idle \\ -1,\ Charge \end{cases} \tag{3}$$

BESS can be defined as

$$E_{BESS,\,t} = \begin{cases} E^{d}_{BESS,\,t},\ Discharging \\ E^{c}_{BESS,\,t},\ Charging \\ \quad 0,\ \ Idle \end{cases} \tag{4}$$

Where $\mathrm{E}^{\mathrm{d}}_{\mathrm{BESS,\,t}}$ and $\mathrm{E}^{\mathrm{c}}_{\mathrm{BESS,\,t}}$ denotes the energy supplied and energy consumed during the discharging and charging states and are expressed as

$$E^{d}_{BESS,t} = -\int_{t'-300}^{t'} P_{BESS}(t)\,dt \tag{5}$$

$$E^{c}_{BESS,t} = \int_{t'-300}^{t'} P_{BESS}(t)\,dt \tag{6}$$

For BESS current (i_{ess}) and rated battery capacity (Ah) state of charge (SOC) can be calculated using (7).

$$SOC_t = 100\times\left(1-\frac{\int i_{ess}dt}{Ah\times 3600}\right) \tag{7}$$

Since the scope of this work is limited to discrete actions by the HEMS agent, the BESS charging and discharging power at any given time is fixed at 1000 Watts and is given by (8). Further, BESS operation is limited by the constraint $20 \le SOC_t \le 100$.

$$|P_{BESS}(t)| = \begin{cases} 1000,\ \text{Charging} \mid \text{Discharging} \\ \quad 0,\ \text{Idle} \end{cases} \tag{8}$$

D. Photovoltaics (PV) System

In this research, the PV system is modeled as a current source and supplies the power based on the look-up table populated by active power values derived from real-world data. Per-unit active power generation across 26 different PV profiles is presented in Fig. 2 and these profiles cover most of the intermittencies. If the power generated by the PV system at any given instance is taken as $P_{PV}(t)$ then PV system can be represented by its energy consumption over five minutes since the last sampling ($t \in [t'-300,\ t']$).

$$E_{PV,\ t} = -\int_{t'-300}^{t'} P_{PV}(t)\,dt \tag{9}$$

E. Grid

The home environment considered in this work is modeled in SIMULINK™ with an infinite source representing the grid. Therefore, it is assumed that the power balance in the system is established. The net energy over the sampling period ($t \in [t'-300,\ t']$) at the point of connection is then given as,

$$E_{net,\ t} = E_{PV,\ t} + E_{Bl,\ t} + E_{BESS,\ t} + \sum_{i=1}^{N} E_{IDA,\ i,\ t} \tag{10}$$

F. Energy cost and Perceived Energy cost

The energy cost for consumption is calculated using the five-minute retail price ($\pi_{C,\ t}$) derived from wholesale prices obtained from the Australian National Electricity Market (NEM). In this research feed-in tariff (FiT) ($\pi_{E,\ t}$) is taken as a fraction of the retail price.

$$\pi_{E,\ t} = \pi_{C,\ t} \times \zeta \tag{11}$$

where ζ is a multiplication factor and is equal to 0.25 in this research. This means for any given instance the FiT is 25% of the retail price.

To ensure consumer preference for appliance operation is taken into consideration by the HEMS agents, perceived net energy and perceived energy cost are introduced. Perceived net energy is given as,

$$\overline{E}_{net,\ t} = E_{PV,\ t} + E_{Bl,\ t} + E_{BESS,\ t} + \sum_{i=1}^{N} (1-\varphi_{IDA,i}) \times E_{IDA,i,\ t} \tag{12}$$

The energy cost and perceived energy cost are derived using (13) and (14) respectively.

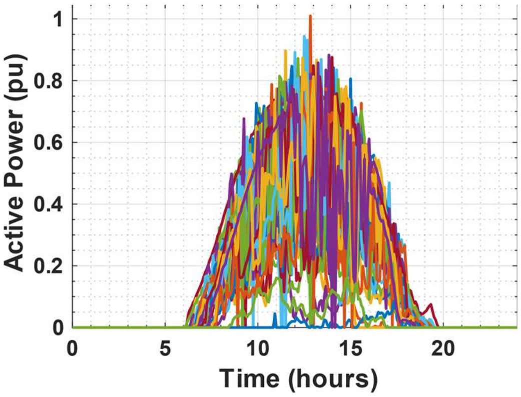

Fig. 2. PV Generation profiles

$$C_{net,\ t} = \begin{cases} \pi_{C,\ t} \times E_{net,\ t}, & if\, E_{net,\ t} \geq 0 \\ \pi_{E,\ t} \times E_{net,\ t}, & if\, E_{net,\ t} < 0 \end{cases} \tag{13}$$

$$\overline{C}_{net,\ t} = \begin{cases} \pi_{C,\ t} \times \overline{E}_{net,\ t}, & if\, \overline{E}_{net,\ t} \geq 0 \\ \pi_{E,\ t} \times \overline{E}_{net,\ t}, & if\, \overline{E}_{net,\ t} < 0 \end{cases} \tag{14}$$

III. MDP Formulation

A tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$ is used to define MDP where $\mathcal{S}$ denotes the state space of the environment, $\mathcal{A}$ denotes the set of actions an agent can take, $\mathcal{P}$ and $\mathcal{R}$ are the transition probability and immediate reward after state transition [24]. Given the objective of the energy management system is to minimize the energy bill while maintaining consumer satisfaction, HEMS can be defined as an optimization problem requiring optimal control of the ID appliance and BESS. The smart home environment can be mathematically expressed using a Markov Decision Process (MDP) framework containing a set of states, actions, and rewards. MDP is also characterized by a Markov property in which only the current state affects the probability distribution of future states.

The same set of observations from the environment is normalized and made available to both agents. The state includes general observations ($\mathcal{S}_G$), base load observations ($\mathcal{S}_{BL}$), observation from IDA ($\mathcal{S}_{IDL}$), BESS ($\mathcal{S}_{BESS}$), and PV ($\mathcal{S}_{PV}$).

$$\mathcal{S}_G = [day,\ t,\ \rho(t)] \tag{15}$$

$$\mathcal{S}_{BL} = \left[E_{Bl,\ t}\right] \tag{16}$$

$$\mathcal{S}_{PV} = \left[E_{PV,\ t}\right] \tag{17}$$

$$\mathcal{S}_{IDL} = \left[E_{IDA,\ t},\ Sch_{On},\ Sch_{Off},\ D,\ \alpha'_{IDA,\ t},\ \varphi,\ \mu_{task}\right] \tag{18}$$

$$\mathcal{S}_{BESS} = \left[E_{BESS,\ t},\ SOC_t\right] \tag{19}$$

where $[Sch_{On},\ Sch_{Off}, D, \varphi]$ represents customers' preferred operation window, the duration of operation, and the preference factor of the IDA for the given day. A customer can have multiple operations of IDA planned on a particular day. μ_{task} denotes task remaining after recent action is taken by the agent.

Each agent in the reinforcement-learning-based multi-agent EMS takes actions to maximize the rewards. Actions and rewards associated with the IDA agent and BESS agent are discussed in the following subsections.

A. IDA – Agent

1) Action space: Since the ID appliance can only be turned "On" or "Off" the actions of IDA-Agent are represented by $\mathcal{A}_{IDL} = [0,1]$.

2) Reward: Rewards for the IDA-Agent can be defined as a function of perceived energy cost ($\overline{C}_{net,\ t}$) incurred by the consumer and the IDA penalty ($\overline{p}_{IDA,\ t}$) associated with the action.

$$\mathcal{R}_{IDA,\ t} = -\left(\overline{C}_{net,\ t} + \overline{p}_{IDA,\ t}\right) \tag{20}$$

where $\bar{p}_{IDA,\ t}$ is given as

$$\bar{\mathrm{p}}_{\mathrm{IDA,\ t}}=\begin{cases} 5,\ \text{if } \alpha'_{\mathrm{IDA,\ t}}=1\ \forall\ \mathrm{t}\notin\left[Sch_{On},\ Sch_{Off}\right] \\ 5,\ \text{if } \alpha'_{\mathrm{IDA,\ t}}=1\ \&\&\ \mu_{\mathrm{task}}=0\ \forall\ t\in\left[Sch_{On},Sch_{Off}\right] \\ 0,\ \text{if } \alpha'_{\mathrm{IDA,\ t}}=1\ \&\&\ \mu_{\mathrm{task}}\neq 0\ \forall\ t\in\left[Sch_{On},Sch_{Off}\right] \\ \varphi\times\mu_{\mathrm{task}},\ \text{if } \alpha'_{\mathrm{IDA,\ t}}=0\ \&\&\ \mu_{\mathrm{task}}\neq 0\ \forall\ t\in\left[Sch_{On},Sch_{Off}\right] \\ 5,\ \text{Otherwise} \end{cases} \tag{21}$$

B. BESS – Agent

1) Action space: As discussed earlier the BESS can be charged, discharged, or left idle. So, the BESS-Agent actions can be represented by the vector $\mathcal{A}_{BESS}$ = [-1, 0,1].

2) Reward: To motivate the BESS-Agent to take actions that maximize overall cost savings, the reward function associated with its actions is defined as

$$\mathcal{R}_{\mathrm{ESS,\ t}} = -\left(\bar{\mathrm{C}}_{\mathrm{net,\ t}} + \bar{\mathrm{p}}_{\mathrm{ESS,\ t}} + \bar{\mathrm{p}}_{\mathrm{IDA,\ t}}\right) \tag{22}$$

where, $\bar{p}_{ESS,\ t}$ is the penalty associated with BESS *SOC* and is defined as

$$\bar{p}_{ESS}=\begin{cases} 5,\ \text{if } SOC\leq 20\ \&\&\ \alpha'_{\mathrm{ESS,\ t}}=1 \\ 5,\ \text{if } SOC\geq 100\ \&\&\ \alpha'_{\mathrm{ESS,\ t}}=-1 \\ 25,\ \text{if } SOC<100\ \&\&\ t=86399 \\ 0,\ \&\text{Otherwise} \end{cases} \tag{23}$$

IV. DRL Algorithm

Deep reinforcement learning (DRL) based HEMS agents use artificial neural networks (ANNs) as function approximators to learn the actions that enable them to achieve their objectives. Considering different environmental constraints, i.e., constraints of the BESS and IDA, and the user preferences, the agents independently learn to take actions to optimize the overall power consumption over several episodes. Each episode is a sequence of timesteps, and the environment rewards the agent at every timestep for actions taken. Depending on the DRL algorithm selected, the agent uses the rewards received from the environment along with the observations and learns to take actions that maximize the rewards it earns.

In this research, two agents, each associated with BESS and IDA are represented by two ANNs. While one ANN called actor or policy network represented by $\pi(A|S;\theta)$ is used to estimate the optimal policy, the other ANN called critic is used to estimate the value function, $\mathrm{V}(S;\emptyset)$. For agents with discrete action set A, for a given state S, the actor with parameters θ estimates the conditional probability of taking each action in set A, and the critic with parameters ∅ outputs the corresponding expected long-term discounted reward. During the training, the DRL agents interact with the environment using a policy randomly selected based on probability distribution and use the states and rewards to optimize the respective actor and critic properties.

For a given mini-batch size of M, the number of actions P, and the divergence limit δ, the TRPO minimizes the actor loss function in (24) subject to constraint in (25), to find the actor parameters.

$$\mathcal{L}_{\mathrm{actor}}(\theta) = -\ \frac{1}{\mathrm{M}}\sum_{\mathrm{i=1}}^{\mathrm{M}}\left(\frac{\pi(\mathrm{A_i}|\mathrm{S_i};\theta)}{\pi(\mathrm{A_i}|\mathrm{S_i};\theta_{\mathrm{old}})}\mathrm{D_i}+\omega\mathcal{H}_{\mathrm{i}}(\theta,\ \mathrm{S_i})\right) \tag{24}$$

$$\frac{1}{\mathrm{M}}\sum_{\mathrm{i=1}}^{\mathrm{M}}\sum_{\mathrm{k=1}}^{\mathrm{P}}\left(\pi(\mathrm{A_k}|\mathrm{S_i};\theta_{\mathrm{old}})\ln\left(\frac{\pi(\mathrm{A_k}|\mathrm{S_i};\theta_{\mathrm{old}})}{\pi(\mathrm{A_k}|\mathrm{S_i};\theta)}\right)\right)\leq\delta \tag{25}$$

where $\mathrm{D_i}$ denotes the advantage function, ω is the entropy loss weight and $\mathcal{H}_{\mathrm{i}}(\theta,\ \mathrm{S_i})$ is the entropy. The TRPO training algorithm using a generalized advantage estimation [23] for the deep reinforcement learning used in this research is presented below.

Algorithm 1 Trust Region Policy Optimization

```
Inputs: Environment states
Initialize actor value function $\pi(S;\theta)$ with random parameters for $\theta$
Initialize critic value function $V(S;\emptyset)$ with random parameters for $\emptyset$
For $episode$ = 1 to $EP$ do
    Follow current actor policy $\pi(S)$ and generate experience consisting of
        $[S_{ts}, A_{ts}, R_{ts+1}, S_{ts+1}, A_{ts+1}, R_{ts+2}, \ldots S_{ts+N-1}, A_{ts+N-1}, R_{ts+N}, S_{ts+N}]$
    Where N is the size of the current experience set and ts represents the starting time step of the experience set.
    For episode step $ts = ts+1, ts+2, ts+3, \ldots, ts+N$ do
        Calculate advantage function $D_t$ using the GAE
            $D_t = \sum_{k=t}^{ts+N-1} (\gamma\lambda)^{k-t}\delta_k$
            $\delta_k = R_t + b\gamma V(S_t;\emptyset)$
        Where $\gamma$ is the discount factor and $\lambda$ is the smoothing factor, and b is given as
            $b=\begin{cases}0, \text{ if } S_{ts+N} \text{ is terminal state}\\ 1, \text{ Otherwise}\end{cases}$
        Calculate the return $G_t$ as
            $G_t = D_t + V(S_t;\emptyset)$
    End for
    Say $\hat{E}$ is the number of epochs, then
    For $e = 1, \hat{E}$
        Randomly sample a mini-batch data set of size M and update the value function parameters after minimizing the value function loss given as
            $\mathcal{L}_{critic}(\emptyset) = \frac{1}{M}\sum_{i=1}^{M}\left(G_i - V(S_i;\emptyset)\right)^2$
        Update the actor parameters after solving (24)
    End for
End for
```

More details on the generalized advantage estimator can be found in [23].

V. Case Studies

A. Experimental Setup

As discussed earlier in this research IDA and BESS are considered. A Washing Machine (WM) with an active power rating of 2.1kW is the IDA considered in this study. Further, a BESS with a capacity of 21.75 Ah and a fixed active power rating of 1 kW is considered in this study. Each episode represents a 24-hour day with DR scheduling starting at 00:00 AM and is mapped to 288 slots of each $\Delta t\ =\ 5$ minutes. In real-world scenarios, the WM may be chosen to operate multiple times a day and each operation may have its customer preference. The operating profiles of the WM used in this research are presented in Table I. The start time and deadline are given in 24-hour format. A maximum of four operations is allowed. When the operation is not planned by the consumer, it can be represented by the tuple (t_{start}, t_{stop}, d, φ) = (26, -1, 0, 0).

TABLE I. WM OPERATIONAL PROFILES WITH PREFERENCES

S.No	Start time (T_{start})	Deadline (T_{stop})	Duration (Hours)	Preference
1[a]	[26, 26, 26, 26]	[-1, -1, -1, -1]	[0, 0, 0, 0]	[0, 0, 0, 0]
2	[15.33, 26, 26, 26]	[18.33, -1, -1, -1]	[1.5, 0, 0, 0]	[0.988, 0, 0, 0]
3	[6.25, 10.38, 26, 26]	[9.25, 13.38, -1, -1]	[2.75, 1.167, 0, 0]	[0.796, 0.679, 0, 0]
4	[6.63, 12.50, 19.51, 26]	[9.63, 15.50, 22.51, -1]	[2.33, 2.50, 1.33, 0]	[0.653, 0.903, 0.197, 0]
5	[6.10, 12.65, 16.35, 19.95]	[9.10, 15.65, 19.35, 22.95]	[2.41, 2, 1.91, 0.50]	[0.480, 0.859, 0.239, 0.168]

[a] WM operation is not preferred.

A PV generation system with an active power rating of 1.5 kW is considered a local energy source and the base load is modeled as a negative power source. The IDA agent consists of an ANN that serves as both actor and critic. The input layer of the policy network of the IDA agent is made up of an input layer with 14 neurons with an activation function of rectified linear units (ReLU) and a fully connected output layer with two neurons representing the on/off actions of the IDA Agent. The ANN representing the value function has 14 neuron input layers followed by five hidden layers with 64, 128, 256, 128, and 64 fully connected neurons with ReLU activation function followed by a single fully connected neuron.

The BESS agent is designed with an actor and critic ANN. The input layer of the policy network of the IDA agent is made up of an input layer with 14 neurons with an activation function of rectified linear units (ReLU) and a fully connected output layer with three neurons representing the charge/discharge/Idle actions of the BESS Agent. The ANN representing the value function has a similar structure to that of the IDA Agent critic. The learning rate used is 0.005.

The home environment and the DRL agents were modeled, trained, and tested using MATLAB/SIMULINK on a computer with Intel Core i7-9700K @ 3.60 GHz. It should be noted that though 8 cores are available parallel processing for multi-agent training is not possible using MATLAB/SIMULINK.

B. Training and Test Datasets

The agents are trained over 18928 episodes and for each episode, a random PV and price profile from a set of 26 one-to-one mapped price and PV profiles, a base-load profile from 7 base-load profiles, and a WM operating profile from the 5 profiles is picked. At the beginning of each episode, the BESS is initialized with a SOC of 40.

The price profiles are derived from the Australian NEM. The original price obtained from NEM is in AUD/MWh and is converted to cents/kWh to meet the real-world retail consumer market scenario. The PV profiles are derived from the real-world data obtained via pvoutput.org for a location near Lismore Heights, NSW Australia. All real-world data used in this research is obtained for all Mondays between October 2021 and March 2022. The base-load profiles used in this research are extracted from the load profiles IEEE European low voltage test feeder which is sampled at 60 s frequency. Three sets of data each having 30 days are used for testing the agents. Each day is made up of a random combination of the PV and price profile, based load profile, and WM profile.

C. Benchmark Methods

The energy cost TRPO-based HEMS system presented in this research is benchmarked against the mixed-integer non-linear programming (MINLP) technique and on-policy policy gradient-based techniques such as vanilla policy gradient and proximal policy optimization.

In the MINLP benchmark, it is assumed that retail electricity price, WM preference, PV generation, and base-load demand are perfectly predicted. The DR scheduling is then formulated as MINLP and solved using a genetic algorithm (GA). Though his benchmark provides optimal results, the presence of randomness makes it impossible to achieve the desired optimum. For PPO and PG benchmarks, agents were trained and tested using the same home environment and data. In the case of the PG agent besides the stochastic policy actor, a baseline critic was used. In both benchmarks, the states, actions, and rewards are kept the same. Further, the ANN structure of the policy network and value function is the same across all the DRL benchmarks.

VI. RESULTS AND DISCUSSION

The average reward obtained by the IDA agent and BESS agent during the training progress is presented in Fig. 3 and Fig. 4 respectively. From these figures, the average reward can be seen increasing rapidly, which eventually begins to converge. However, for the BESS agent, the average rewards can be seen raising and dropping over a small window while the average reward for IDA Agent mostly remains stable.

The comparison of the total energy cost for the test sets of TRPO-based HEMS with other benchmarks discussed in section V is presented in Fig. 5. As can be seen from Fig. 5 the optimal energy cost is set by the MINLP optimization using GA. The TRPO based HEMS performs better than PPO based HEMS in two out of three test sets and better than PG based HEMS in all three test sets.

VII. CONCLUSION

In this article, a model-free on-policy TRPO-based multi-agent customer-centric home energy management with an entropy-loss weight of 0.01 is studied considering the uncertainty of the electricity prices, PV generation, and base load. A washing machine (WM) with multiple operational profiles is considered an IDA. Each operation profile has a customer preference which affects the decision of HEMS. The proposed HEMS exercises discrete control of the IDA and BESS in a decentralized manner. Real-world data is used to train and validate this HEMS system.

Further, the performance in terms of the energy cost of the TRPO-based multi-agent HEMS is compared with the other on-policy approaches such as PPO and PG, while the MINLP optimization using GA is used to set the optimal performance expectations for HEMS. Using three sets of test data it is demonstrated that the TRPO multi-agent HEMS achieves near-optimal energy cost.

In this study, the problem appears simple due to the use of discrete action space. However, this work provides optimistic results for the extension of TRPO HEMS into the continuous action space. In the discrete action space, it takes approximately 62 hours to train the agents over 18928 episodes. In this context, it is recommended that for real-time application of TRPO HEMS in the continuous action space, the agents are trained offline.

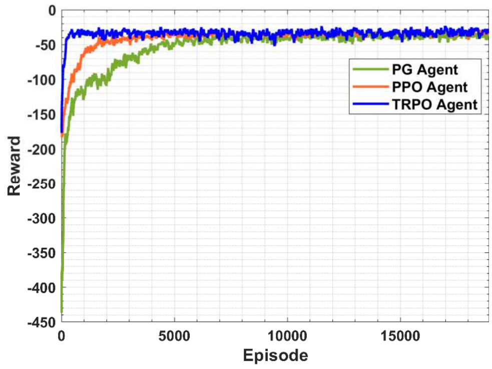

Fig. 3. IDA Agent Training Progress

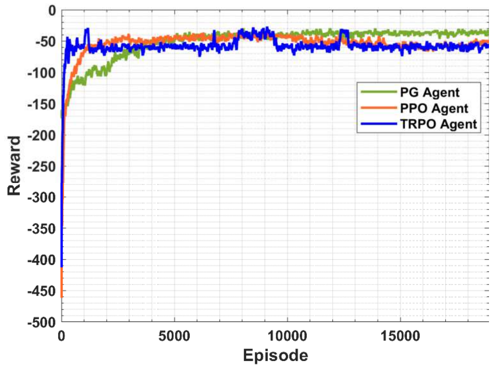

Fig. 4. BESS Agent Training Progress

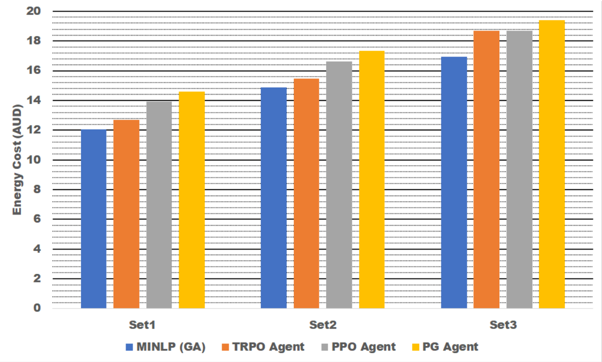

Fig. 5. Performance comparison of HEMS Techniques

References

[1] J. Schoene, V. Zheglov, D. Houseman, J. C. Smith, and A. Ellis, “Photovoltaics in distribution systems - Integration issues and simulation challenges,” in *IEEE Power and Energy Society General Meeting*, 2013, pp. 1–5, doi: 10.1109/PESMG.2013.6672879.

[2] “Annual Energy Outlook 2022 - U.S. Energy Information Administration (EIA),” 2022. https://www.eia.gov/outlooks/aeo/narrative/electricity/sub-topic-01.php (accessed Dec. 27, 2022).

[3] D. Zhang, S. Li, M. Sun, and Z. O’Neill, “An Optimal and Learning-Based Demand Response and Home Energy Management System,” *IEEE Trans. Smart Grid*, vol. 7, no. 4, pp. 1790–1801, Jul. 2016, doi: 10.1109/TSG.2016.2552169.

[4] H. T. Nguyen, D. T. Nguyen, and L. B. Le, “Energy management for households with solar assisted thermal load considering renewable energy and price uncertainty,” *IEEE Trans. Smart Grid*, vol. 6, no. 1, pp. 301–314, 2015, doi: 10.1109/TSG.2014.2350831.

[5] M. Beaudin and H. Zareipour, “Home energy management systems: A review of modelling and complexity,” *Renew. Sustain. Energy Rev.*, vol. 45, pp. 318–335, May 2015, doi: 10.1016/j.rser.2015.01.046.

[6] I. Gomes, K. Bot, M. G. Ruano, and A. Ruano, “Recent Techniques Used in Home Energy Management Systems: A Review,” *Energies*, vol. 15, no. 8, p. 2866, Apr. 2022, doi: 10.3390/en15082866.

[7] J. Lyu, T. Ye, M. Xu, G. Ma, Y. Wang, and M. Li, “Price-sensitive home energy management method based on Pareto optimisation,” *Int. J. Sustain. Eng.*, vol. 14, no. 3, pp. 433–441, May 2021, doi: 10.1080/19397038.2020.1822948.

[8] H. T. Dinh and D. Kim, “An Optimal Energy-Saving Home Energy Management Supporting User Comfort and Electricity Selling With Different Prices,” *IEEE Access*, vol. 9, pp. 9235–9249, 2021, doi: 10.1109/ACCESS.2021.3050757.

[9] B. Jeddi, Y. Mishra, and G. Ledwich, “Differential Dynamic Programming Based Home Energy Management Scheduler,” *IEEE Trans. Sustain. Energy*, vol. 11, no. 3, pp. 1427–1437, Jul. 2020, doi: 10.1109/TSTE.2019.2927237.

[10] Z. Zheng, Z. Sun, J. Pan, and X. Luo, “An integrated smart home energy management model based on a pyramid taxonomy for residential houses with photovoltaic-battery systems,” *Appl. Energy*, vol. 298, no. June, p. 117159, 2021, doi: 10.1016/j.apenergy.2021.117159.

[11] H. Saberi, C. Zhang, and Z. Y. Dong, “Data-Driven Distributionally Robust Hierarchical Coordination for Home Energy Management,” *IEEE Trans. Smart Grid*, vol. 12, no. 5, pp. 4090–4101, 2021, doi: 10.1109/TSG.2021.3088433.

[12] R. Lu, S. H. Hong, and M. Yu, “Demand Response for Home Energy Management Using Reinforcement Learning and Artificial Neural Network,” *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6629–6639, Nov. 2019, doi: 10.1109/TSG.2019.2909266.

[13] J. R. Vázquez-Canteli and Z. Nagy, “Reinforcement learning for demand response: A review of algorithms and modeling techniques,” *Appl. Energy*, vol. 235, no. April 2018, pp. 1072–1089, Feb. 2019, doi: 10.1016/j.apenergy.2018.11.002.

[14] A. Tittaferrante and A. Yassine, “Multiadvisor Reinforcement Learning for Multiagent Multiobjective Smart Home Energy Control,” *IEEE Trans. Artif. Intell.*, vol. 3, no. 4, pp. 581–594, Aug. 2022, doi: 10.1109/TAI.2021.3125918.

[15] T. Yang, L. Zhao, W. Li, J. Wu, and A. Y. Zomaya, “Towards healthy and cost-effective indoor environment management in smart homes: A deep reinforcement learning approach,” *Appl. Energy*, vol. 300, no. July, p. 117335, Oct. 2021, doi: 10.1016/j.apenergy.2021.117335.

[16] I. Zenginis, J. Vardakas, N. E. Koltsaklis, and C. Verikoukis, “Smart Home’s Energy Management Through a Clustering-Based Reinforcement Learning Approach,” *IEEE Internet Things J.*, vol. 9, no. 17, pp. 16363–16371, 2022, doi: 10.1109/JIOT.2022.3152586.

[17] Y. Ye, D. Qiu, H. Wang, Y. Tang, and G. Strbac, “Real-Time Autonomous Residential Demand Response Management Based on Twin Delayed Deep Deterministic Policy Gradient Learning,” *Energies*, vol. 14, no. 3, p. 531, Jan. 2021, doi: 10.3390/en14030531.

[18] S. Lee and D.-H. Choi, “Federated Reinforcement Learning for Energy Management of Multiple Smart Homes With Distributed Energy Resources,” *IEEE Trans. Ind. Informatics*, vol. 18, no. 1, pp. 488–497, Jan. 2022, doi: 10.1109/TII.2020.3035451.

[19] X. Zhang *et al.*, “Two-Stage Reinforcement Learning Policy Search for Grid-Interactive Building Control,” *IEEE Trans. Smart Grid*, vol. 13, no. 3, pp. 1976–1987, May 2022, doi: 10.1109/TSG.2022.3141625.

[20] H. Li, Z. Wan, and H. He, “Real-Time Residential Demand Response,” *IEEE Trans. Smart Grid*, vol. 11, no. 5, pp. 4144–4154, Sep. 2020, doi: 10.1109/TSG.2020.2978061.

[21] O. Al-Ani and S. Das, “Reinforcement Learning: Theory and Applications in HEMS,” *Energies*, vol. 15, no. 17, p. 6392, Sep. 2022, doi: 10.3390/en15176392.

[22] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, “Trust Region Policy Optimization,” *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 3, pp. 1889–1897, Feb. 2015, [Online]. Available: http://arxiv.org/abs/1502.05477.

[23] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, “High-Dimensional Continuous Control Using Generalized Advantage Estimation,” *4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.*, pp. 1–14, Jun. 2015, [Online]. Available: http://arxiv.org/abs/1506.02438.

[24] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, “Deep reinforcement learning: A brief survey,” *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017, doi: 10.1109/MSP.2017.2743240.