Large-Scale Home Energy Management Using Entropy-Based Collective Multiagent Deep Reinforcement Learning Framework

Yaodong Yang[1] , **Jianye Hao**[1*] , **Yan Zheng**[1] and **Chao Yu**[2]
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China
yydapple@gmail.com, jianye.hao@tju.edu.cn, yanzheng@tju.edu.cn, cy496@dlut.edu.cn

Abstract

Smart grids are contributing to the demand-side management by integrating electronic equipment, distributed energy generation and storage, and advanced meters and controllers. With the increasing adoption of electric vehicles and distributed energy generation and storage systems, residential energy management is drawing more and more attention, which is regarded as being critical to demand-supply balancing and peak load reduction. In this paper, we focus on a microgrid scenario in which modern homes interact together under a large-scale setting to better optimize their electricity cost. We first incentivize households to form a group using an economic stimulus. Then we formulate the energy expense optimization problem of the household community as a multi-agent coordination problem and present an Entropy-Based Collective Multiagent Deep Reinforcement Learning (EB-C-MADRL) framework to address it. Experiments with various real-world data demonstrate that EB-C-MADRL can reduce both the long-term group power consumption cost and daily peak demand effectively compared with existing approaches.

1 Introduction

Meeting the growing energy demand due to the presence of more volatile types of loads raises a major challenge for the power grid [Robu *et al.*, 2013]. In order to satisfy demand that varies sharply, companies usually have to install additional generation capacity to meet the peak demand with a heavy price. The requirement of massive investment into peaking power generation, in turn, results in higher costs for end-users. At the same time, distributed renewable generation such as wind and solar energy is gaining prominence and is perceived as vital to achieving cost and carbon reduction [Shweta Jain *et al.*, 2014]. However, renewable generation has the unreliable nature that it is quite intermittent and sensitive to weather changes. As a result, the increase in electricity supply from renewable sources leads to larger fluctuations which makes the power grid hard to maintain the demand-supply balance. With the increasing number of active electricity customers and the advent of decentralized power generation technologies, the peak load and supply-demand imbalance have received more and more attention by energy generation and distribution companies [Zhang *et al.*, 2015].

To handle the above problems, the demand-side management (DSM) [Strbac, 2008] has been proposed, which aims to adjust the consumer's energy activities such as shifting customers' consumption from peak hours to off-peak hours. From the perspective of energy providers, lots of DSM technologies have been developed to reschedule consumption such as Time-of-Use (TOU) tariffs [Shweta Jain *et al.*, 2014]. On the other hand, there are also many DSM techniques focusing on the home energy management, such as dynamic programming [Yuan-Yih Hsu and Chung-Ching Su, 1991], game theory [Mohsenian-Rad *et al.*, 2010] and reinforcement learning (RL) [O'Neill *et al.*, 2010]. However, these works only consider different subsets of the home power systems instead of the complete home energy system. Another disadvantage of these solutions is the rigid schedule for end users' appliances usage. Recently, smart homes combined with the distributed energy generation (DG) and distributed energy storage (DS) show the great possibility for the revolution of the power grid [Palensky and Dietrich, 2011; Logenthiran *et al.*, 2016]. It provides us with opportunities of unfreezing the rigid schedule for users with the emerging DG and DS. RL based DSM techniques for the smart home was first investigated in [Berlink and Costa, 2015] and then extended in [Wu *et al.*, 2018] with electric vehicles (EV). Compared with traditional methods, RL could learn flexible energy management strategies and well suits the promising smart home scenario without manual adjustments.

However, these smart home DSM works focus on optimizing the energy activities for a single household. They do not consider the aggregate effect of all customers shifting the demand to low price/low demand periods, which would result in a sudden increase in demand and overloads on the transformer [Dusparic *et al.*, 2013]. On the contrary, one exception is that Dusparic et al. [Dusparic *et al.*, 2013] proposed a multiagent approach to manage the energy demand of a small house group. Their strategy, called W-learning, consists of integrating independent Q-learning agents that one agent schedules on the appliance usage for one home. However, the role

[*] Corresponding author: Jianye Hao.

of households considered in their approach is the traditional power consumer and the rigid appliance usage schedules violate user habits and bring inconvenience. To the best of our knowledge, no work has investigated DSM for the concernful large-scale smart-home energy management problem.

To this end, in this paper, we research on the user-friendly DSM techniques for a smart home community. We first entice the community with an incentive mechanism to form a trading group and consider this scenario as partially observable Markov Games. Then we propose an entropy-based collective multiagent reinforcement learning (MARL) framework to address the large-scale energy cost optimization problem. Our framework optimizes the energy consumption by abstracting the market dynamics to mitigate the non-stationary problem and avoid the action space explosion. Besides, it utilizes the concept of entropy to reduce peak load. Experiments simulated from real-world data exhibit excellent performance of our framework compared with previous approaches.

The remainder of this paper is organized as follows. We introduce the Markov Games, the smart home and deep reinforcement learning algorithms in Section 2. Then we describe the microgrid electricity market in Section 3. And in Section 4 we explain our EB-C-MADRL framework in details. Finally, we demonstrate the effectiveness of our framework in the microgrid simulated from real-world data in Section 5. Conclusions and the future work are provided in Section 6.

2 Background

2.1 Markov Games

In this work, we consider a multiagent extension of the Markov decision process (MDP) called partially observable Markov games [Littman, 1994]. A Markov game for n agents identified by $i \in I \equiv \{1,2,...,n\}$ is defined by a set of states S describing the global state, a set of actions $A_1,...,A_n$ and a set of observations $O_1,...,O_n$ for each agent. Each agent i chooses actions according to its policy $\pi_i : O_i \times A_i \to [0,1]$, which produces the next state according to the transition function $T : S \times A_1 \times ... \times A_n \to S$. After the transition, each agent i obtains its reward $r_i(s,a_1,...,a_n) : S \times A_1 \times ... \times A_n \times I \to R$, and receives a private observation correlated with the state $o_i : S \to O_i$. The initial states are determined by distribution $\rho : S \to [0,1]$. Each agent i aims to maximize its own total expected return $J_i = \mathbb{E}_{a^i \sim \pi_1,...,a^N \sim \pi_N, s \sim T} \sum_{t=0}^{T} \gamma^t r_{i,t}(s,a_1,...,a_n)$ where γ is a discount factor and T is the time horizon.

2.2 Smart Home

We follow the smart home concept in [Wu *et al.*, 2018] as shown in Figure 1. The main components can be divided into:

- base load electricity consumption, consumed from conventional household appliances;
- EV charging power consumption;
- micro-generation, such as Photovoltaics (PV);
- energy storage, the home battery for storing power;
- and home energy management system (EMS) for DSM.

Advanced meters are installed at home, enabling bi-directional communication. A bi-directional flow guarantees smart homes acquire or insert energy into the power grid.

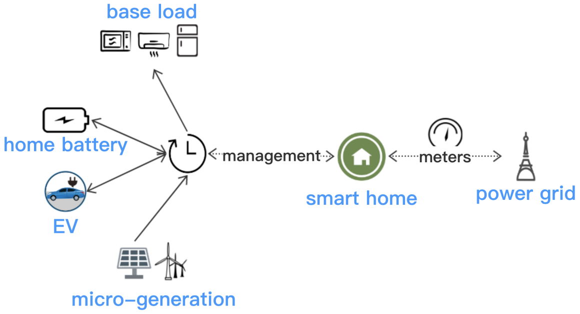

Figure 1: An illustration of a smart home.

2.3 Deep Reinforcement Learning

Here we introduce Deep Q-Networks (DQN) and Advantage Actor-Critic (A2C). For maximizing the accumulated expected return J, Q-learning uses an action-value function for policy π as $Q^{\pi}(s,a) = \mathbb{E}[J|s_t = s, a_t = a]$ and updates its Q-values based on each experience given by $(s,a,s^{'},r)$:

$$Q(s,a) \longleftarrow Q(s,a) + \alpha * [r + \gamma * Q(s^{'},a^{'}) - Q(s,a)], \quad (1)$$

where α is the learning rate and γ is the discount factor. The ε-greedy is used to trade off exploration-exploitation.

Deep Q-Networks (DQN)

DQN represents the action-value function with a deep neural network parameterized by θ. The Q-function can be recursively rewritten as $Q^{\pi}(s_t,a_t) = \mathbb{E}_{s_{t+1}}[r_{t+1} + \gamma \mathbb{E}_{a_{t+1} \sim \pi}[Q^{\pi}(s_{t+1},a_{t+1})]]$ for updating. DQNs use a replay buffer to store the transition $(s,a,s^{'},r)$. DQN is trained with experience replay by minimizing the squared TD-error:

$$\mathscr{L}(\theta) = \sum_k [(y_k^{DQN} - Q(s_t,a_t;\theta))^2], \quad (2)$$

where $y_k^{DQN} = r_t + \gamma \max_{a_{t+1}} Q(s_{t+1},a_{t+1};\theta^-)$. θ^- are the target network parameters.

Advantage Actor-Critic (A2C)

Policy gradient techniques aim to estimate the gradient of expected returns with respect to the parameters of its policy:

$$\nabla_{\theta} J(\pi_{\theta}) = \mathbb{E}_{a \sim \pi_{\theta}}[\nabla_{\theta} log(\pi_{\theta}(a_t|s_t)) \sum_{t^{'}=t}^{\infty} \gamma^{t^{'}-t} r_{t^{'}}], \quad (3)$$

where $J(\pi_{\theta})$ is the accumulated expected return. $\sum_{t^{'}=t}^{\infty} \gamma^{t^{'}-t} r_{t^{'}}$ term leads to high variance, as the return can vary drastically between episodes. Actor-Critic methods aim to ameliorate this issue by using a function approximation of the expected return $V^{\pi}(s) = \mathbb{E}_{a \sim \pi(s)}[J|s_t = s, a_t = a]$ and replacing the original return in the policy gradient estimator with this function. A2C employs two deep networks, a policy network to learn policy, $\pi(a|s;\theta)$ parameterized by θ and a value network to learn value function, $V^{\pi}(s;\varphi)$ parameterized by φ.

The policy network is updated according to the policy loss:

$$\begin{aligned} \nabla_{\theta} J(\pi_{\theta}) &= \mathbb{E}_{a \sim \pi_{\theta}}[\nabla_{\theta} log(\pi_{\theta}(a_t|s_t))A(s_t,a_t)], \\ where\ A(s_t,a_t) &= Q(s_t,a_t) - \sum_a \pi_{\theta}(a|s_t)Q(s_t,a) \\ &= r_t + \gamma * V^{\pi_{\theta}}(s_{t+1};\varphi) - V^{\pi_{\theta}}(s_t;\varphi). \end{aligned} \quad (4)$$

3 Microgrid Electricity Market

3.1 Microgrid Dynamics

We model the smart home in Section 2.2 as the basic unit, which mainly follows settings in [Wu *et al.*, 2018]. The tariff scheme in the microgrid is TOU, which gives different prices according to the time slot. At the beginning of each time slot, given the price signal, all households need to schedule the power trading and EV charging. To avoid the rigid schedule for the appliances usage, the base load from traditional appliances activities remains unchanged. The microgrid market can be modeled as partially observable Markov Games ($\boldsymbol{G}$) and below we present each component of $\boldsymbol{G}$.

States The state for each home is composed of the following local observations: electricity price p_t at current time slot t, home based battery state of charge $H_{b,t}$, home based PV energy generation amount $H_{p,t}$, base load power consumption $H_{l,t}$, EV charging availability $E_{a,t}$, time to EV departure $E_{d,t}$ and current EV battery state of charge $E_{b,t}$. $H_{b,t}$ and $E_{b,t}$ are 100% when the battery is fully charged and 0% when fully discharged. $E_{a,t}$ is the charging availability of EV (set as 1 when available and 0 otherwise). $E_{d,t}$ shows how many hours remains before EV departure. Formally, household i's status $o_t^i \in O$ is defined as $o_t^i = (p_t, H_{b,t}^i, H_{p,t}^i, H_{l,t}^i, E_{a,t}^i, E_{d,t}^i, E_{b,t}^i)$.

Actions At the beginning of each time slot t, the home EMS needs to decide two actions: $P_{c,t}$ for power trading amount and $C_{e,t}$ for the EV charging rate. Homes can sell energy to the power grid and purchase energy for consumption and storage. Both the EV and home battery can be charged or discharged with continuous values (from zero to the maximal allowed charging rate). The home battery is responsible for redundancy caused by $P_{c,t}$ and $C_{e,t}$ with satisfying Equation 5.

$$H_{p,t} + P_{c,t} = C_{e,t} + C_{b,t} + H_{l,t}, \tag{5}$$

where $C_{b,t}$ is the charging rate for the home battery. Following [Wu *et al.*, 2018], the EV charging action $C_{e,t}$ is designed as five discrete rates: 100% charging rate, 50% charging rate, 0, 50% discharging rate and 100% discharging rate. Each household's original action space for power trading contains a continuous value domain, which ranges from the sum of its current stored and generation energy to the maximum energy the household can consume during the current time slot. To reduce the power trading action space, we set two sub-action sets for trading power. For individuality, each household's action values refer to its yesterday's power consumption and generation statistics. The home EMS purchases power at different rates based on the average hourly gross power consumption σ of previous day. Home EMS can also sell power at different rates based on the average hourly gross power generation β of previous day. Thus, the action set for $P_{c,t}$ is $\{\sigma, 75\%\sigma, 50\%\sigma, 25\%\sigma, 0, -25\%\beta, -50\%\beta, -75\%\beta, -\beta\}$. Action 0 means no trading or no charging. Let B_e denotes the EV battery capacity and B_h the home battery capacity. When charging or discharging, the efficiency η is set at 0.9. At each time slot, we check the availability of EV to decide whether applying a constraint to set $C_{e,t}$ as 0. After determining $C_{e,t}$ and $P_{c,t}$, the power surplus or shortage is calculated as $C_t = H_{p,t} + H_{b,t} \cdot B_h - H_{l,t} - C_{e,t}$. The home battery is updated passively for redundant operation if $C_t \neq 0$ after performing $C_{e,t}$ and $P_{c,t}$. The battery charging rate and discharging rate are no more than the maximal charging power rate while deciding the allowed actions.

Rewards The electricity cost is calculated as follows:

$$r_t = \begin{cases} P_{c,t} * p_b, & if\ P_{c,t} \geq 0, \\ P_{c,t} * p_s, & if\ P_{c,t} < 0 \end{cases}, \tag{6}$$

where p_b and p_s are the current electricity buying price and selling price separately. p_s is smaller than p_b because of specific utility programs such as long-distance transmission fees.

Transitions At each time slot, after all households deciding actions, the microgrid processes these energy decisions to deterministically step into the next state. All the decisions determine electricity prices as shown in the next Section 3.2. After energy actions done, each household updates its state based on its own energy actions and the TOU price signal.

3.2 Group Incentive Mechanism

In the local microgrid, each household can interact with power grid directly. But it is better for these households to coordinate into a group as in [Cailliere *et al.*, 2016] to locally match the production and the consumption of the group, which not only helps balance the demand and supply for the power grid but also saves costs for households. We design an incentive-driven market mechanism to attract users to join in a group for coordination. The microgrid market mechanism has two trading processes: the internal trading process and the external trading process. Households trade inside the group first to satisfy the demand of each other. If the internal trading cannot fully meet the group, then the external smart grid will deal with the unsatisfied demand. To encourage users to actively participate in such a microgrid, we set the internal power price and external power prices as follows:

$$p_{os,t} \leq p_{in,t} \leq p_{ob,t}, \tag{7}$$

where $p_{os,t}$ and $p_{ob,t}$ are the current power selling and buying prices for customers, which are different due to utility program fees. $p_{in,t}$ is the current internal power trading price. With such a constraint price, households are willing to trade inside first because of the better price and the same usage guarantee as directly interacting with the outer grid. On the other hand, smart grid may benefit from such a market mechanism because of possible peak load reduction by the group coordination. It is common to see extra aggregate demand or supply after internal trading. Thus, the final cleaning price for electricity integrated with external trading is:

$$\begin{aligned} p_{s,t} &= \begin{cases} \frac{p_{in,t}\psi_{b,t} + p_{os,t}(\psi_{s,t} - \psi_{b,t})}{\psi_{s,t}}, & if\, \psi_{s,t} \geq \psi_{b,t} \\ p_{in,t}, & if\, \psi_{s,t} < \psi_{b,t} \end{cases} \\ p_{b,t} &= \begin{cases} p_{in,t}, & if\, \psi_{s,t} \geq \psi_{b,t} \\ \frac{p_{in,t}\psi_{s,t} + p_{ob,t}(\psi_{b,t} - \psi_{s,t})}{\psi_{b,t}}, & if\, \psi_{s,t} < \psi_{b,t} \end{cases}, \end{aligned} \tag{8}$$

where $p_{s,t}$ and $p_{b,t}$ are the power selling price and buying price at time t. $\psi_{s,t}$ and $\psi_{b,t}$ are the total power selling and buying amount at t. We can observe that $p_{s,t}$ is always in the range $[p_{os,t}, p_{in,t}]$ and $p_{b,t}$ is always in $[p_{in,t}, p_{ob,t}]$, which indicates customers are willing to trading inside for better prices.

For reducing energy cost, households need capture overall market dynamics to decide when to buy and sell power. Through the incentive mechanism, we turn the smart home community a multiagent system, where each agent's reward is determined by trading prices affected through the aggregation of the group behavior. Promoting the group coordination is needed for better cost optimization and such a problem is inherently multiagent and can be solved by MARL approaches.

4 Entropy-Based Collective Multiagent Reinforcement Learning Framework

4.1 EB-C-MARL Algorithm

Previous works of the home energy management mainly focus on optimizing a single household's or a small-scale customers' electricity operating cost [Atzeni *et al.*, 2013; Berlink and Costa, 2015; Wu *et al.*, 2018]. As they are never designed for large-scale multiagent systems, applying existing approaches directly for the household community may bring the potential harm such as raising a new peak load and even jeopardizes the infrastructure of power grids. Thus, it is necessary to investigate the home EMS algorithm from the perspective of a household community as presented in this paper. Algorithm 1 describes our EB-C-MADRL framework.

Algorithm 1 Entropy-based collective MARL in microgrid

```
Input: the episode number $M$; each episode's step number $T$; the household number $N$.
Output: constructed DQN or A2C.
 1: Initialize the deep network with random weights;
 2: Set the initial collective group behavior approximations $a'_{s,t} = 0, a'_{b,t} = 0, \vec{C}'_{e,t} = \vec{0}, t = 1,2,..,24$;
 3: for $episode = 1,2,...,M$ do
 4:     Initialize state $s_1$ and all households' state $(o_1^1, o_1^2, ..., o_1^N)$;
 5:     for $t = 1,2,...,T$ do
 6:         for $i = 1,2,...,N$ do
 7:             Household $i$ chooses $P_{c,t}^i$ and $C_{e,t}^i$ based on its state $o_t^i$ and the approximations of market dynamics $a'_{s,t}$, $a'_{b,t}$ and $\vec{C}'_{e,t}$;
 8:         end for
 9:         Process actions, collect market information $a_{s,t}$, $a_{b,t}$ and $\vec{C}_{e,t}$, update $a'_{s,t\%24} = a_{s,t}$, $a'_{b,t\%24} = a_{b,t}$ and $\vec{C}_{e,t\%24} = \vec{C}_{e,t}$;
10:         for $i = 1,2,...,N$ do
11:             Compute individual entropy $h_t^i$ by Equation 15 and electricity operating cost $r_t^i$ by Equation 17;
12:         end for
13:         Observe $s_{t+1}$ and all households' state $(o_t^1, o_t^2, ..., o_t^N)$;
14:         Update neural networks with Equation 2 or Equation 4;
15:     end for
16: end for
```

The descriptions of the process of EB-C-MADRL are shown in Algorithm 1. Line 1 initializes network parameters. Line 2 initializes approximations of the collective group behavior for 24 hours. Line 4 initializes the starting state and households' observations. Line 6-8 chooses actions for each household based on its observation and the approximations of collective group behavior detailed in Section 4.2. Line 9 executes actions and abstracts market dynamics. Line 10-13 shows that once actions are performed, each household's reward and entropy are calculated as in Section 4.3 and the next state is observed. Line 14 updates the neural network.

4.2 Collective Group Behavior

All households are frequently interacting in the local microgrid market as each household determines its action choices at each time slot. Such a massive dynamic property raises huge challenges for the home EMS algorithms particularly RL based ones. One primary problem is that each agent's policy is changing as training progresses, and the environment becomes non-stationary from the perspective of any individual agent [Lowe *et al.*, 2017]. Even if we could obtain actions from other agents without the privacy conservation to mitigate the non-stationary issue, in the large-scale multiagent systems, the representation of the joint action becomes another prominent problem. The joint action space of the agents grows exponentially with the number of agents, which makes the value function learning extremely hard [Yang *et al.*, 2018]. The mean field approach in [Yang *et al.*, 2018] is not applicable in microgrid since there is no "neighbor" for agents to interact with and each agent instead interact with the microgrid market only. However, in market settings where agents are influenced from their collective action effect, we could represent such collective influence by the market dynamics abstraction to avoid the action space explosion and finally mitigate the non-stationary problem. To approximate the collective influence of agents, we also resort to the daily periodicity in the power grid. Next we describe how the market dynamics abstraction is integrated with DQN and A2C.

Collective DQN

As Section 3.2 shows, households can be divided into buyer and seller groups dynamically to determine the electricity buying and selling prices. From each agent i's perspective, it is coordinating with the mircogird market instead of directly interacting with any individual agent. Thus, we could abstract market macro-actions to replace other agents' joint action to simplify the multiagent Q-function significantly. The joint-action Q-function can be simplified as follows:

$$Q^i(s, a^1, a^2, ..., a^N) \equiv Q^i(s, a^i, a^{market}), \tag{9}$$

where the abstraction of market dynamics a^{market} includes the seller group collective action a_s (total supply amount), the buyer group collective action a_b (total demand amount) and group EV charging distribution $\vec{C}_e$. One additional privacy benefit is that each household only need to access to its own states and abstract market dynamics and cannot access to any other household. Then we obtain Equation 10:

$$Q^i(s, a^i, a^{market}) \approx Q^i(o^i, a^i, a_s, a_b, \vec{C}_e). \tag{10}$$

The abstractions of current market dynamics cannot be exactly obtained as all households make decisions at the same time. Instead we propose using group collective actions at the same time slot in the previous day to approximate current market dynamics. As inhabitants exhibit similar living habits in daily life, it is reasonable to assume their energy activities have similar daily periodicity especially for group statistics:

$$Q^i(o^i, a^i, a_s, a_b, \vec{C}_e) \approx Q^i(o^i, a^i, a_s^{'}, a_b^{'}, \vec{C}_e^{'}), \tag{11}$$

where $a_s^{'}$, $a_b^{'}$ and $\vec{C}_e^{'}$ are group action statistics at one day ago. The loss function for collective DQN is defined as:

$$\mathscr{L}(\theta) = \sum_k [(y_k^{DQN} - Q^i(o_t^i, a_t^i, a_{s,t}^{'}, a_{b,t}^{'}, \vec{C}_{e,t}^{'}; \theta))^2], \tag{12}$$

where $y_k^{DQN} = r_t^i + \gamma \max_{a_{t+1}^i} Q^i(o_{t+1}^i, a_{t+1}^i, a_{s,t+1}^{'}, a_{b,t+1}^{'}, \vec{C}_{e,t+1}^{'}; \theta^-)$.

Collective A2C

Similarly, the collective market actions also benefit each agent's home EMS policy when applying A2C by providing abstract information of other agents' current policies.

$$\begin{aligned} \pi^i(s, a^1, ..., a^{i-1}, a^{i+1}, ..., a^N) &\equiv \pi^i(s, a^{market}) \\ \approx \pi^i(o^i, a_s, a_b, \vec{C}_{e,t}) &\approx \pi^i(o^i, a_s^{'}, a_b^{'}, \vec{C}_{e,t}^{'}). \end{aligned} \tag{13}$$

With $\tilde{o}^i = (o^i, a_s^{'}, a_b^{'}, \vec{C}_e^{'})$ for simple presentation, the loss of policy network of collective A2C is given by:

$$\begin{aligned} \nabla_\theta J(\pi_\theta^i) &= \mathbb{E}_{a \sim \pi_\theta^i}[\nabla_\theta log(\pi_\theta^i(a_t^i|\tilde{o}_t^i))A(\tilde{o}_t^i, a_t^i)], \\ where\, A(\tilde{o}^i, a_t^i) &= Q^i(\tilde{o}_t^i, a_t^i) - \sum_{a^i} \pi_\theta^i(a^i|\tilde{o}_t^i) Q^i(\tilde{o}_t^i, a^i) \\ &= (r_t^i + \gamma * V^{\pi_\theta^i}(\tilde{o}_{t+1}^i; \varphi)) - V^{\pi_\theta^i}(\tilde{o}_t^i; \varphi). \end{aligned} \tag{14}$$

$V^{\pi_\theta^i}(\tilde{o}_t^i; \varphi) = \sum_{a^i} \pi_\theta^i(a^i|\tilde{o}_t^i) Q^i(\tilde{o}_t^i, a^i)$ denotes the expected return.

4.3 Reward Shaping with Individual Entropy

For reducing the daily peak load, we use a mechanism called **individual entropy** to diversify household EV charging to different time slots. As [Muratori, 2018] shows, with more market share, the uncoordinated EV charging would result in higher peak loads as EV charging usually happens when people arrive home after work. Even learning with RL algorithms for each household, the uncoordinated learning will result in high peak load. It is because that EV would charge in the low electricity price period coincidentally and there is no explicit factor to diversify the EV charging in training. Inspired by [Verma *et al.*, 2018] which maximizes agent density entropy to make taxi drivers well-proportioned in different regions, we utilize the entropy to diversify the EV charging behavior.

Unlike using the system's total entropy for each agent in [Verma *et al.*, 2018], we use more accurate individual entropy h_t^i to assign credits of contributing to the system's entropy for each household. Intuitively, if one household chooses a low-frequency action different from others, a higher bonus would be assigned to the household as it contributes more to the system's entropy H_t. If user i chooses an action a_t^i from action set A at t, then h_t^i for user i is calculated as follows:

$$h_t^i = \frac{-\log p_{a_t^i}}{N}, \tag{15}$$

where p_{a_t} is the frequency of action a_t in all actions performed at t. h_t^i gives the incentive to choose a different action from current high-frequency actions. Therefore, it helps reduce the peak load by mitigating the phenomenon that households charge EV concurrently. The individual entropy is accurate credit assignment of the system's entropy which represents the distribution degree of EV charging behavior:

$$\sum_i h_t^i = \sum_i \frac{-\log p_{a_t^i}}{N} = \sum_{a_t^i} \frac{-n_{a_t^i} \log p_{a_t^i}}{N} = \sum_{a_t^i} -p_{a_t^i} \log p_{a_t^i} = H_t. \tag{16}$$

Also, the individual entropy is hoped to only affect the EV charging action which often happens in low-price periods. Therefore, we add an adjustment term to lower the weight of entropy when the current TOU price $p_{b,t}$ is high. The final reward for household i with individual entropy is given as:

$$r_t^i = \begin{cases} P_{c,t}^i * p_{b,t} + \frac{coef}{p_{b,t}} * h_t^i, & if P_{c,t}^i \geq 0, \\ P_{c,t}^i * p_{s,t} + \frac{coef}{p_{b,t}} * h_t^i, & if P_{c,t}^i < 0, \end{cases} \tag{17}$$

where $p_{b,t}$ and $p_{s,t}$ are the power buying and selling prices at t. $P_{c,t}^i$ and $P_{c,t}^i$ are household i's current power purchasing and selling amount. h_t^i is the entropy bonus term for household i in Equation 15 while $coef$ is the entropy coefficient. Each agent receives the reward signal including h_t^i from the market.

5 Experiments and Analysis

5.1 Experiment Settings

We use various real-world data to establish our microgrid simulation environment. TOU electricity price from [Wu *et al.*, 2018] and residential power consumption data in 2013 from [Muratori, 2018] are used. The power consumption data contains 200 households' consumption patterns in US. The home battery and EV configurations are the same as in [Wu *et al.*, 2018]. The daily driving distance data used to calculate the EV status when arriving home obeys *gamma distribution* [Lin *et al.*, 2012] with shape at 1.6 and scale at 20. The parameters are highly frequent in daily vehicle driving distance datasets [Lin *et al.*, 2012] and the average daily driving distance is close to [Plötz *et al.*, 2017] in the US. We use irradiance data in 2013 of Building 8167, Gatton Campus [of Queensland, 2018] for PV generation. For user diversity, we vary each household's installed power capacity from $1.1kWp$ to $4.4kWp$ at uniform distribution. For simplification, we set the electricity selling price in the external trading process p_{os} at half of current TOU price, which is used as external electricity purchase price p_{ob} for households. The internal trading price is set as the average between current p_{os} and p_{ob}. Next, we validate EB-C-MADRL in the simulated microgrid.

5.2 Validating the Collective Group Behavior

We first validate the collective group behavior abstraction. At each time t, agents i in total 200 agents receives $(p_t, H_{b,t}^i, H_{p,t}^i, H_{l,t}^i, E_{a,t}^i, E_{d,t}^i, E_{b,t}^i, a_{s,t}^{'}, a_{b,t}^{'}, \vec{C}_{e,t}^{'})$. We adopt DQN and A2C as control algorithms for each household. For speeding up the training process in the large-scale settings, we consider parameter sharing in the home EMS strategies as households can be viewed as homogeneous agents and the learned policy is a generalized strategy for households.

Such a training paradigm is widely adopted in multiagent reinforcement learning algorithms and shows excellent performance [Yang *et al.*, 2018]. We use the first four weeks' data in 2013 for training and the remaining data for evaluation. The training phase contains 125 episodes and each episode lasts for 28 days with interaction interval being an hour.

We compare the performance of the proposed control algorithms with two baselines: rule-based and DQN, which performs best in the previous single household case [Wu *et al.*, 2018]. The rule-based control algorithm is called *Naive-greedy* policy described in [Berlink and Costa, 2015], which charges the EV when arriving home and sell the energy when there is a power surplus. Such a policy results in high electricity operation costs as EVs are charged at high-price periods. Then we augment both DQN and A2C with market dynamics approximations to validate the collective group behavior abstraction. Table 1 shows the electricity operating results. And Daily Peak Load is calculated by adding all daily peak loads up in testing, which is not optimized temporarily.

Algorithm	Operating Cost ($)	Internal Trading (kWh)	Peak Load (kwh)
Naive Greedy	-263195±0.06%	107277±0.01%	453303±0.06%
DQN	-111133±17.0%	198762±14.06%	421048±3.14%
A2C	-92174±3.39%	225251±2.70%	478322±2.57%
Collective DQN	-93087±13.40%	188949±6.92%	429021±3.47%
Collective A2C	-88878±4.47%	219969±3.38%	465816±3.10%

Table 1: Group Operating Results with 95% Confidence Interval

All results are averaged over the running of 20 random seeds. *Naive-Greedy* policy performs poorly as it charges EV in high-price periods and never utilizes the home battery to store energy for the later use. With the collective group behavior approximations, collective DQN performs better than DQN by 16.24% in terms of the total operating cost while collective A2C's cost is less than A2C by 3.58%. The collective A2C achieves the least operating cost by more internal trading and shifting the EV charging to off-peak hours.

5.3 Validating the Individual Entropy Mechanism

Despite achieving the least cost, collective A2C still has high peak loads caused by the uncoordinated EV charging aggregation. To be specific, RL agents would independently learn to optimize cost by discharging EV when arriving home to make profits and charging EV in low-price periods, which centrally results in high peak loads. To mitigate the new peaks, we enhance collective DQN and collective A2C with individual entropy in Equation 17 to encourage agents to diversify EV charging. Table 2 gives the results of related methods and the EB-C-MADRL framework with the best settings. Compared with DQN, entropy-based collective A2C (EB-C-A2C) achieves 24.69% cost reduction and 5.15% peak load reduction. Note that, the operation cost is counted based on the real electricity cost and does not include the entropy term.

Next we investigate EB-C-A2C and entropy-based collective DQN (EB-C-DQN) using different coefficients as Figure 2. EB-C-A2C achieves the best results for both the operating cost and peak load. EB-C-DQN seems not sensitive to the individual entropy because of its deterministic policy. On the other hand, collective A2C's stochastic policy can change its action more easily with the slight adjustment of action probabilities to get the higher entropy bonus. With the entropy coefficient being 1, the group daily peak load could achieve the least amount while the electricity operating cost reduces a lot dramatically. Such a cost reduction mainly results from two reasons. One reason is that the individual entropy implicitly divides action space into two sub-spaces by directly affecting one action sub-space only, leading to easier separate learning of power trading and EV charging actions. Another reason is that the combination of the approximation of current EV charging distribution and the individual entropy reflects microgrid dynamics and other households' policy tendency, which further alleviates the non-stationary problem. Too large entropy coefficient would increase both the group peak load and electricity operating cost. This is because that households learn to discharge and charge EV repeatedly to get more entropy bonus which makes energy cost ignored.

Algorithm	Operating Cost ($)	Peak Load (kwh)
Naive Greedy	-263195.44±168.58	453302.63±282.11
DQN	-111133.42±18897.06	421048.18±13241.82
Collective A2C	-88878.34±3971.80	465816.24±14448.64
EB-C-A2C	-83689.13±1098.20	399381.48±11503.41

Table 2: Group Operating Results with 95% Confidence Interval

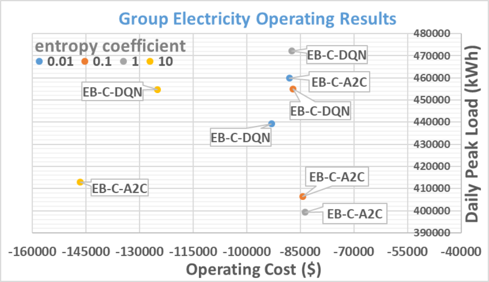

Figure 2: Group power operating results with individual entropy.

6 Conclusion and Future Work

In this paper, we focus on a large-scale smart home EMS problem. First, we model the group energy optimization as a multiagent coordination problem in an incentive market. Then we propose EB-C-MADRL to learn home EMS control policies. Experiments exhibit superior performance of our method for operating cost saving and peak load reduction.

As future work, auction mechanisms could be considered to make microgrid market more liberal. Also, using EB-C-MARL integrated with more advanced DRL algorithms to explore other large-scale multiagent markets such as e-commerce markets or water allocation markets is interesting.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (Grant Nos.: 61702362, U1836214), Special Program of Artificial Intelligence and Special Program of Artificial Intelligence of Tianjin Municipal Science and Technology Commission (No.: 569 17ZXRGGX00150).

References

[Atzeni *et al.*, 2013] Italo Atzeni, Luis Garcia Ordóñez, Gesualdo Scutari, Daniel Pérez Palomar, and Javier Rodríguez Fonollosa. Demand-side management via distributed energy generation and storage optimization. *IEEE Transactions on Smart Grid*, 4:866–876, 2013.

[Berlink and Costa, 2015] Heider Berlink and Anna Helena Reali Costa. Batch reinforcement learning for smart home energy management. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 2561–2567, 2015.

[Cailliere *et al.*, 2016] Romain François Cailliere, Samir Aknine, and Antoine Nongaillard. Multi-agent mechanism for efficient cooperative use of energy. In *Proceedings of the 15th International Conference on Autonomous Agents & Multiagent Systems*, pages 1365–1366, 2016.

[Dusparic *et al.*, 2013] Ivana Dusparic, Colin Harris, Andrei Marinescu, Vinny Cahill, and Siobhan Clarke. Multi-agent residential demand response based on load forecasting. In *Proceedings of the 1st IEEE Conference on Technologies for Sustainability (SusTech)*, pages 90–96, 2013.

[Lin *et al.*, 2012] Zhenhong Lin, Jing Dong, Changzheng Liu, and David Greene. Estimation of energy use by plug-in hybrid electric vehicles: Validating gamma distribution for representing random daily driving distance. *Transportation Research Record: Journal of the Transportation Research Board*, 2287(1):37–43, 2012.

[Littman, 1994] Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings*, pages 157–163. Elsevier, 1994.

[Logenthiran *et al.*, 2016] Thillainathan Logenthiran, W. Li, and W. L. Woo. Intelligent multi-agent system for smart home energy management. In *Proceedings of the IEEE Innovative Smart Grid Technologies - Asia*, pages 1–6, 2016.

[Lowe *et al.*, 2017] Ryan Lowe, YI WU, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31th Advances in Neural Information Processing Systems*, pages 6379–6390, 2017.

[Mohsenian-Rad *et al.*, 2010] Amir-Hamed Mohsenian-Rad, Vincent W. S. Wong, Juri Jatskevich, Robert Schober, and Alberto Leon-Garcia. Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid. *IEEE Transactions on Smart Grid*, 1(3):320–331, 2010.

[Muratori, 2018] Matteo Muratori. Impact of uncoordinated plug-in electric vehicle charging on residential power demand. *Nature Energy*, 3(3):193–201, 2018.

[of Queensland, 2018] University of Queensland. Uq solar photovoltaic data. http://solar.uq.edu.au/user/reportPower.php, 2018.

[O'Neill *et al.*, 2010] Daniel O'Neill, Marco Levorato, Andrea Goldsmith, and Urbashi Mitra. Residential demand response using reinforcement learning. In *Proceedings of the 1st IEEE International Conference on Smart Grid Communications*, pages 409–414, 2010.

[Palensky and Dietrich, 2011] Peter Palensky and Dietmar Dietrich. Demand side management: Demand response, intelligent energy systems, and smart loads. *IEEE Transactions on Industrial Informatics*, 7(3):381–388, 2011.

[Plötz *et al.*, 2017] Patrick Plötz, Niklas Jakobsson, and Frances Sprei. On the distribution of individual daily driving distances. *Transportation Research Part B: Methodological*, 101:213–227, 2017.

[Robu *et al.*, 2013] Valentin Robu, Enrico Gerding, Sebastian Stein, David C. Parkes, Alex Rogers, and Nicholas R. Jennings. An online mechanism for multi-unit demand and its application to plug-in hybrid electric vehicle charging. *Journal of Artificial Intelligence Research*, 48:175–230, 2013.

[Shweta Jain *et al.*, 2014] Shweta Jain, Balakrishnan Narayanaswamy, and Y. Narahari. A multiarmed bandit incentive mechanism for crowdsourcing demand response in smart grids. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 721–727, 2014.

[Strbac, 2008] Goran Strbac. Demand side management: Benefits and challenges. *Energy Policy*, 36(12):4419–4426, 2008.

[Verma *et al.*, 2018] Tanvi Verma, Pradeep Varakantham, and Hoong Chuin Lau. Entropy controlled non-stationarity for improving performance of independent learners in anonymous MARL settings. *arXiv preprint*, abs/1803.09928, 2018.

[Wu *et al.*, 2018] Di Wu, Rabusseau Guillaume, François lavet Vincent, Precup Doina, and Boulet Benoit. Optimizing home energy management and electric vehicle charging with reinforcement learning. In *Proceedings of the 16th Adaptive Learning Agents*, 2018.

[Yang *et al.*, 2018] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5567–5576. PMLR, 2018.

[Yuan-Yih Hsu and Chung-Ching Su, 1991] Yuan-Yih Hsu and Chung-Ching Su. Dispatch of direct load control using dynamic programming. *IEEE Transactions on Power Systems*, 6(3):1056–1061, 1991.

[Zhang *et al.*, 2015] Baosen Zhang, Ramesh Johari, and Ram Rajagopal. Competition and coalition formation of renewable power producers. *IEEE Transactions on Power Systems*, 30(3):1624–1632, 2015.