Privacy Preserving Load Control of Residential Microgrid via Deep Reinforcement Learning

Zhaoming Qin, Di Liu, Haochen Hua, *Member IEEE*, and Junwei Cao, *Senior Member, IEEE*

***Abstract*—Demand side management has been proved to be effective in improving the operating efficiency of microgrids, while posing a severe threat to user privacy. This paper proposes a novel privacy preserving load control scheme for the residential microgrid, in which the microgrid operator manages a multitude of home appliances including electric vehicles (EVs) and air conditioners (ACs). This problem is formulated as a partially observable Markov decision process, since users' privacy information including indoor temperatures associated with ACs and arrival/departure times of EVs cannot be observed by microgrid operator. To address the formulated problem with high-dimensional continuous action space caused by massive controllable appliances, we develop a novel deep reinforcement learning algorithm by introducing credit assignment mechanism. Moreover, we integrate recurrent neural network to accommodate the partial observability of state due to privacy issues. Simulation results demonstrate the superiority and flexibility of the developed algorithm and verify the advantages of the proposed scheme compared with prior privacy preserving load control method.**

***Index Terms*—Deep reinforcement learning, load control, privacy preserving, residential microgrid.**

I. Introduction

A. Background and Motivation

With the development of informatization and intellectualization, traditional electric power system has evolved from a top-down and relatively static structure to a bottom-up and active one [1], [2]. As a typical load at the end of the energy transmission chain, residential loads evidently have great potential for energy management and regulation, accounting for nearly 40% of the energy consumption and CO_2 emission in some developed countries [3]. It is notable that the advanced energy management, storage and sharing technologies make it possible to realize the self-balance of power in a microgrid equipped with high proportional renewable energy sources (RESs) [4]. Particularly, due to the non-adjustability of RESs, the energy management on the demand side plays a vital role in the power balance of the microgrid.

However, the load control in residential microgrid relies on the fact that the user information is attributable and fine-grained, raising the privacy concerns. For example, to improve the user's thermal comfort, the microgrid operator controls the air conditioner (AC) depending on the user's indoor temperature information [5]-[9]. Similarly, in order to schedule electric vehicle (EV) charging, the information including the arrival and departure times of EVs is assumed to be available [10]-[13], such that user travel information is predictable and user privacy is at a risk of leakage. Furthermore, most existing load control methods for residential microgrid are designed on the assumption that the prior knowledge of user behavioral preference is available to the microgrid operator [14]-[16], which poses a severe threat to user privacy. One case is that the user's temperature preference can be inferred from the user's thermal comfort loss function. Therefore, it is essential to address the privacy concerns of residential load control.

This work was sponsored in part by Tsinghua-Toyota Joint Research Institute Cross-discipline Program, and in part by Fundamental Research Funds for the Central Universities of China, Grant B200201071. *(Corresponding author: Junwei Cao.)*

Z. Qin and D. Liu are with the Department of Automation, Tsinghua University, Beijing, 100084, P. R. China.

H. Hua is with the College of Energy and Electrical Engineering, Hohai University, Nanjing, 211100, P. R. China.

J. Cao is with Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, 100084, P. R. China (email: jcao@tsinghua.edu.cn).

B. Literature Review

Although tremendous research outputs have been dedicated to developing the direct load control of residential microgrid [5]-[16], few of these studies focuses on the privacy issues [17]. This can be attributed to two main reasons. First, most existing privacy preserving methods focus on incentive-based [18] or price-based [19] programs to address the privacy concerns incurred by advanced metering infrastructure, while ignoring the privacy threats posed by direct load control. Second, it is intractable for conventional load control methods to achieve desirable control effect when private information cannot be observed.

Fortunately, by combining reinforcement learning (RL) with deep learning, the deep reinforcement learning (DRL) technologies have brought a new solution to this challenge. First, the model-free RL can cope with the load control problems without prior knowledge and explicit models [8]. Second, benefiting from the powerful feature extraction capability of deep neural networks, DRL algorithms have the potential to learn a good policy given very limited information in the privacy preserving environment [20].

In recent years, the value-based DRL algorithms, which estimate the action-value function of RL with deep neural networks, have been successfully applied to the field of direct load control. The deep Q-network (DQN) is used for EV charging scheduling in [21], while dueling DQN is applied to optimize the demand response management of interruptible load in [22]. However, the value-based DRL can only solve the problems with discrete and low-dimensional action spaces, since it relies on maximizing the action-value function over the whole action space, which in the continuous case requires a complicated optimization process at every decision step [23].

To resolve the tasks with continuous action spaces, policy-based DRL algorithms directly generate the specific value of action (deterministic policy) or the probability distribution of

action (stochastic policy). For example, an energy management scheme for AC control based on deep deterministic policy gradient (DDPG) algorithm is proposed in [8], and asynchronous advantage actor-critic (A3C) is applied to handle the economic dispatch problem of virtual power plant in [24].

C. Scope and Contribution

In this paper, we propose a privacy preserving load control scheme to minimize the operation cost of the whole microgrid while considering the user experience. To be specific, the indoor temperatures associated with ACs are fully preserved. Furthermore, the arrival and departure times of EVs would not be provided in advance. Then, to accommodate the partial observability involved by the privacy issues, the considered control problem is described as a partially observable Markov decision process (POMDP), rather than the general Markov decision process (MDP). We address the formulated problem based on a policy-based DRL algorithm called advantage actor-critic (A2C) algorithm which performs better on control problems with continuous action spaces compared to the value-based DRL algorithms (e.g., DQN and its variants), and possesses better exploration capability owing to its stochastic policy representation compared to DDPG [25]. Moreover, to tackle the high-dimensional action space and the partial observability of state, we develop the vectorized advantage actor-critic (VA2C) algorithm and integrate recurrent neural network (RNN). Finally, numerical simulations are used to demonstrate the superior performance of our developed algorithm over the benchmark DRL algorithm and to verify the advantages of proposed scheme compared to the existing privacy preserving scheme.

The importance and contribution of this paper can be summarized as follows.

1) A privacy preserving load control scheme considering user experience is proposed. In contrast with existing AC control schemes [5]-[9], in this work, the indoor temperatures and thermal comfort functions are completely preserved, which eliminates the possibility of corresponding privacy leakage. Moreover, compared with previous works involved in EV charging scheduling [10]-[13], the arrival and departure times of EVs are not required to be known in advance, which further enhances the protection for user privacy.

2) POMDP is applied to formulate the privacy preserving load control problem, in which the long-term profit of microgrid operator and user experience are considered simultaneously. Compared with standard MDP settings adopted in most DRL-based works, e.g., [21], [22], [24]-[26], POMDP provides a framework for accommodating privacy issues. Accordingly, an integrated deep neural network, comprising a RNN architecture and multi-layer perceptron (MLP), is employed to extract information from limited observation.

3) We adopt a typical model-free policy-based DRL algorithm to resolve the load control problem without prior knowledge of user behavior and model dynamics. Moreover, we introduce the credit assignment mechanism to A2C, developing the vectorized A2C (VA2C) algorithm. The developed algorithm can reduce the variance of estimate and stabilize the learning when handling the problems with high dimensional action spaces.

4) Simulation results demonstrate the superiority of the developed VA2C algorithm over the benchmark A2C algorithm. Compared to existing privacy preserving scheme, the proposed scheme provides better user experience with less energy consumption and operation cost. Moreover, the flexibility and scalability of VA2C algorithm is verified.

II. Problem Formulation

A. System Description

As shown in Fig. 1, the considered load control scenario is composed of a microgrid operator, smart homes, distributed generators (DGs) and battery energy storage devices (BESs). We suppose that the microgrid operator operates in discrete time, i.e., $t \in \mathcal{T} = \{0,1,\dots,T-1\}$, where T is the time horizon. The duration of a time slot is denoted by Δt. At each time slot t, the microgrid operator collects information from smart homes and other energy devices, and then makes continuous control signals with the aim of minimizing the operation cost of the whole microgrid while improving the user experience.

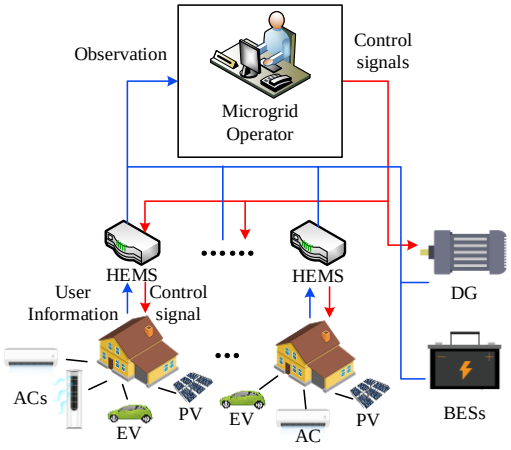

Fig. 1. System structure of the considered load control scheme.

In the considered microgrid, each smart home is equipped with EVs, ACs, photovaltaic panels (PVs) and non-adjustable loads. Assuming that the numbers of EVs and ACs in the microgrid are M and N, respectively, the set of EVs and ACs are denoted by $\{1,\dots,M\}$ and $\{1,\dots,N\}$, respectively. For simplicity, the non-adjustable loads in all smart homes are considered as a whole and are called base loads. Similarly, all PVs are treated as one non-adjustable power generation device.

It is designed that each smart home is controlled by a home energy management system (HEMS). As the gateway of each smart home, the HEMS is desigenated to upload very limited information of the smart home to microgrid operator, directly control adjustable appliances (EVs and ACs) according to the control signals from microgrid operator, collect the user comfort loss and feed it back to microgrid operator. For the purpose of protecting user privacy, some crucial information, such as room temperature, arrival and departure time for EVs, should not be directly exposed to microgrid operator. In this sense, the uploaded user information only includes current

charging demands of EVs and current usage status of ACs.

B. Partial Observable Markov Decision Process Formulation

Since microgrid operator cannot observe all the information of the system, the energy mangement problem is formulated as a POMDP, rather than MDP. A general MDP can be described as a 4-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$, where $\mathcal{S}, \mathcal{A}, \mathcal{P}, r$ are the state space, action space, transition dynamics and reward function, respectively. At time slot t, an agent observes a state $\boldsymbol{s}_t$ from state space $\mathcal{S}$, and selects an action $\boldsymbol{a}_t$ from action space $\mathcal{A}$. After performing action $\boldsymbol{a}_t$, state $\boldsymbol{s}_t$ transitions to state $\boldsymbol{s}_{t+1}$ with probability distribution $\mathcal{P}(\boldsymbol{s}_t, \boldsymbol{a}_t)$. Additionally, the agent receives a scalar reward $r_t = r(\boldsymbol{s}_t, \boldsymbol{a}_t)$.

Similarly, the POMDP can be described as a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \Omega)$ where Ω is the observation space [27]. At time slot t, the agent can only observe $\boldsymbol{o}_t$ from observation space Ω, rather than the full state $\boldsymbol{s}_t$. Let EV_i denote the i-th EV and AC_j denote the j-th AC, the detail of POMDP formulation is presented as follows.

1) *State*: The state $\boldsymbol{s}_t$ contains the full information of considered microgrid at time slot t, including the total output power of PVs represented by P_t^{PV}, the total power of base loads represented by P_t^{BL}, the output power of DGs represented by P_t^{DG}, the state of charge (SOC) of BESs represented by SOC_t, the outdoor temperature of smart homes represented by T_t^{out}, the charging demands of EVs represented by vector $\boldsymbol{E}_t^{EV} = \left[E_t^{EV_1}, \dots, E_t^{EV_M}\right]$, the indoor temperatures controlled by ACs represented by vector $\boldsymbol{T}_t^{AC} = \left[T_t^{AC_1}, \dots, T_t^{AC_N}\right]$, and usage status of ACs represented by vector $\boldsymbol{S}_t^{AC} = \left[S_t^{AC_1}, \dots, S_t^{AC_N}\right]$ where the binary variable $S_t^{AC_j}$ indicates whether AC_j is used by corresponding occupant at time slot t. Note that the ACs can be controlled by microgrid operator at any time, even if some occupants are not using ACs. In this sense, the state $\boldsymbol{s}_t$ is expressed as follows,

$$\boldsymbol{s}_t = [P_t^{PV}, P_t^{BL}, T_t^{out}, P_t^{DG}, SOC_t, \boldsymbol{E}_t^{EV}, \boldsymbol{T}_t^{AC}, \boldsymbol{S}_t^{AC}]. \quad (1)$$

2) *Observation*: In most existing work on direct load control, the indoor temperatures are used by default as the indispensable information to decide the control signals of ACs [20]. Similarly, the EV charging scheduling relies on the information about arrival and departure times for EVs. However, for privacy purposes, all the above-mentioned private information is kept locally and cannot be obtained by microgrid operator. Therefore, the observation $\boldsymbol{o}_t$ is defined as follows,

$$\boldsymbol{o}_t = [P_t^{PV}, P_t^{BL}, T_t^{out}, P_t^{DG}, SOC_t, \boldsymbol{E}_t^{EV}, \boldsymbol{S}_t^{AC}]. \quad (2)$$

3) *Action*: In the considered microgrid, the controllable appliances include DGs, M EVs and N ACs. The action $\boldsymbol{a}_t$ in continuous action space $\mathcal{A}$ is defined as

$$\boldsymbol{a}_t = \left[u_t^{DG}, P_t^{EV_1}, \dots, P_t^{EV_M}, P_t^{AC_1}, \dots, P_t^{AC_N}\right], \quad (3)$$

where u_t^{DG}, $P_t^{EV_i}$ and $P_t^{AC_j}$ are the control signal of DGs, the power of EV_i and AC_j, respectively. Furthermore, u_t^{DG}, $P_t^{EV_i}$ and $P_t^{AC_j}$ are supposed to be restricted in $[0,1]$, $\left[0, P_{max}^{EV_i}\right]$ and $\left[0, P_{max}^{AC_j}\right]$,where $P_{max}^{EV_i}$ and $P_{max}^{AC_j}$ denote the maximum charging power of EV_i and maximum input power of AC_j, respectively. We denote the dimension of $\boldsymbol{a}_t$ by K. Here, $K = 1 + M + N$.

4) *Transition dynamics*: With the state $\boldsymbol{s}_t$ and action $\boldsymbol{a}_t$, we intend to describe the dynamics of state as following general forms.

$P_t^{PV}, P_t^{BL}, T_t^{out}$: Influenced by many unknown factors, the power of PVs, the power of base loads and outdoor temperature are difficult to be modeled precisely. In this work, the real historical data is directly employed, including power and temperature data [28], [29].

P_t^{DG}: The dynamics of output power of DGs can be modeled as

$$P_{t+1}^{DG} = f^{DG}(P_t^{DG}, u_t^{DG}), \quad (4)$$

where $f^{DG}(\cdot,\cdot)$ is the transition function of power of DGs with respect to the output power and control signal of DGs.

SOC_t: The dynamics of SOC is described as

$$SOC_{t+1} = f^{BES}(SOC_t, P_t^{BES}), \quad (5)$$

where $f^{BES}(\cdot,\cdot)$ is the transition function of SOC with respect to SOC and charging/discharging power of BESs. Moreover, P_t^{BES} is determined by the following power balance constraint (6), rather than directly controlled by the microgrid operator.

$$P_t^{BES} = \sum_{i=1}^{M} P_t^{EV_i} + \sum_{j=1}^{N} P_t^{AC_j} + P_t^{BL} - P_t^{PV} - P_t^{DG}. \quad (6)$$

$E_t^{EV_i}$: The transition dynamics of the charging demand of EV_i throughout the time horizon are written as

$$E_{t+1}^{EV_i} = \begin{cases} E_{init}^{EV_i}, & \text{if} \quad t+1 = t_a^{EV_i} \\ f^{EV_i}\left(E_t^{EV_i}, P_t^{EV_i}\right), & \text{if} \quad t_a^{EV_i} < t+1 < t_d^{EV_i} \\ 0, & \text{otherwise} \end{cases}, \quad (7)$$

where $E_{init}^{EV_i}$, $t_a^{EV_i}$ and $t_d^{EV_i}$ represent the initial charging demand, arrival time and departure time of EV_i. Under our proposed privacy preserving scheme, any information of these random variables is unknown, in contrast with some existing load control schemes [10]-[13]. The first line in (7) denotes the transition of $E_t^{EV_i}$ when EV_i is arriving, while the second line represents the transition of $E_t^{EV_i}$ during the charging process of EV_i, in which $f^{EV_i}(\cdot,\cdot)$ is the transition function with respect to charging demand and charging power of EV_i.

$T_t^{AC_j}$: The indoor temperature is affected by factors including AC power and outdoor temperature, the general dynamics of which is described as follows

$$T_{t+1}^{AC_j} = f^{AC_j}\left(T_t^{AC_j}, T_t^{out}, P_t^{AC_j}\right), \quad (8)$$

where $f^{AC_j}(\cdot,\cdot,\cdot)$ is the transition function of indoor temperature attributed to AC_j with respect to indoor temperature, outdoor temperature and input power of AC_j.

$S_t^{AC_j}$: The transition of $S_t^{AC_j}$ reflects the occupant's usage behavior for AC_j, determined by two random variables as follows

$$S_{t+1}^{AC_j} = \begin{cases} 1, & \text{if } t_e^{AC_j} \le t+1 < t_l^{AC_j} \\ 0, & \text{otherwise} \end{cases}. \quad (9)$$

Here, $t_e^{AC_j}$ and $t_l^{AC_j}$ denote the time when the occupant enters and leaves the room where AC_j is located, respectively. For privacy reasons, the information about $t_e^{AC_j}$ and $t_l^{AC_j}$ cannot be captured by microgrid operator, which is similar to the random variables motivating the dynamics of $E_t^{EV_i}$.

5) *Reward*: The target of our proposed load control scheme is to minimize the operation cost of the whole microgrid while ensuring the user experience. In this sense, the reward at time slot t is established as follows,

$$r_t = -\left(C_t^{DG} + C_t^{BES} + \sum_{i=1}^{M} C_t^{EV_i} + \sum_{j=1}^{N} C_t^{AC_j}\right). \quad (11)$$

In (11), C_t^{DG} represents the power generation cost of DGs at time slot t, defined as

$$C_t^{DG} = g^{DG}(P_t^{DG})\Delta t, \quad (12)$$

where $g^{DG}(\cdot)$ denotes the unit time generation cost function with respect to output power of DGs. C_t^{BES} denotes the degradation cost of BESs at time slot t. For the purpose of BESs lifetime extension, it is desired to reduce the depth of discharge/charge power. Therefore, a general function is employed to calculate C_t^{BES} as follows

$$C_t^{BES} = g^{BES}(SOC_t, P_t^{BES}), \quad (13)$$

where $g^{BES}(\cdot,\cdot)$ is the degradation cost function with respect to SOC and charging/discharging power.

The user dissatisfaction incurred by the charging scheduling of EV_i is denoted by $C_t^{EV_i}$, determined by the owner of EV_i. Since the user dissatisfaction only occurs when EV is not well charged at the departure time, the user dissatisfaction of EV charging is indeed a sparse reward [30] and can be represented as follows,

$$C_t^{EV_i} = \begin{cases} g^{EV_i}\left(E_t^{EV_i}\right), & \text{if } t = t_d^{EV_i} \\ 0, & \text{if } t \ne t_d^{EV_i}, \end{cases} \quad (14)$$

where $g^{EV_i}(\cdot)$ is the user comfort loss function associated with unsatisfied charging demand of EV_i. Note that the departure time $t_d^{EV_i}$ is unavailable to microgrid operator. In (11), $C_t^{AC_j}$ denotes the thermal comfort loss caused by room temperature $T_t^{AC_j}$, which occurs when AC_j is used by the corresponding occupant. Therefore, $C_t^{AC_j}$ is presented as follows,

$$C_t^{AC_j} = \begin{cases} g^{AC_j}\left(T_t^{AC_j}\right), & \text{if } S_t^{AC_j} = 1 \\ 0, & \text{if } S_t^{AC_j} = 0 \end{cases}, \quad (15)$$

where $g^{AC_j}(\cdot)$ is the user thermal comfort loss function associated with indoor temperature $T_t^{AC_j}$.

It is worth mentioning that there are no restrictive assumptions on the forms of $g^{EV_i}(\cdot)$ and $g^{AC_j}(\cdot)$ which are completed determined by corresponding users.

III. Solution via Deep Reinforcement Learning

The policy-based DRL algorithms have been achieved remarkable performance on control problems with continuous action spaces [25]-[26]. However, there exist some obstacles to straightforwardly employing these algorithms to handle the formulated problem. First, it is intractable to explore the high-dimensional continuous action space, caused by a large multitude of controllable appliances (EVs and ACs). Second, these algorithms rely on the capability to perceive complete information at each time slot [20], while the microgrid operator suffers from partial observability of state.

In this section, we first establish the basic framework for solving the formulated POMDP problem with RL. Then, A2C, a benchmark DRL algorithm to address control problems with continuous action space, is intuitively introduced. Finally, we elaborate the developed VA2C algorithm and explain how the developed algorithm cope with the proposed privacy preserving load control scheme.

A. Reinforcement Learning with POMDP

In Section II, we describe the privacy preserving load control problem as POMDP, leading to the corresponding RL settings as follows.

Consider a RL problem where an agent interacts with an environment $\mathcal{E}$ over a series of discrete time slots. At each time slot t, the agent obtains an observation $\boldsymbol{o}_t$ of state $\boldsymbol{s}_t$ and chooses an action $\boldsymbol{a}_t$ from the action space $\mathcal{A}$ according to its policy π, where $\pi: \Omega \to \mathcal{A}$ is a mapping from observation $\boldsymbol{o}_t$ to actions $\boldsymbol{a}_t$. In return, the agent receives a scalar reward r_t and the observation $\boldsymbol{o}_{t+1}$ of next state $\boldsymbol{s}_{t+1}$. The process continues until the terminal slot T.

The discounted cumulative reward from time slot t is defined as

$$R_t = \sum_{\tau=t}^{T-1} \gamma^{\tau-t} r_\tau, \quad (16)$$

where $\gamma \in (0,1]$ is the discount factor. The goal in RL is to learn a policy π which maximizes the expected return from the start distribution, i.e.,

$$J[\pi] = \mathbb{E}_{\boldsymbol{a}_t \sim \pi}[R_0]. \quad (17)$$

Here, J is the objective function of RL and $\mathbb{E}$ is mathematical expectation.

B. Advantage Actor-Critic

We formulate π as a stochastic policy $\pi(\cdot \mid \cdot; \theta): \mathcal{A} \times \Omega \to \mathbb{R}$ parameterized by θ. The policy $\pi(\boldsymbol{a}_t|\boldsymbol{o}_t;\theta)$ indicates the probability for performing action $\boldsymbol{a}_t \in \mathcal{A}$ given observation $\boldsymbol{o}_t \in \Omega$. To measure the performance of policy π given a specific observation, the value function $V^\pi(\cdot): \Omega \to \mathbb{R}$ is defined as

$$V^\pi(\boldsymbol{o}_t) = \mathbb{E}_{\boldsymbol{a}_{\tau \ge t} \sim \pi}[R_t|\boldsymbol{o}_t], \quad (18)$$

which is the expected return in observation $\boldsymbol{o}_t$ and following policy π.

The vanilla policy-based algorithms [31] update the parameter θ in the direction of

$$\nabla_\theta J[\pi(\cdot \mid \cdot;\theta)] = \mathbb{E}[\sum_{t=0}^{T-1}(\nabla_\theta \log \pi(\boldsymbol{a}_t|\boldsymbol{o}_t;\theta)\, R_t)], \quad (19)$$

where ∇_θ denotes the gradient to θ. The classic Monte Carlo method estimates this gradient by random samples, which introduces high variability and causes unstable learning. To reduce the variance of this estimate and to increase stability, one way is to subtract the cumulated reward R_t by a baseline $V(\boldsymbol{o}_t;\theta_v)$ which is an estimate of the value function $V^\pi(\boldsymbol{o}_t)$. Combined with the fact that the expectation of R_t can be presented as $\mathbb{E}[r_t + \gamma V^\pi(o_{t+1})]$, the gradient in (19) can be approximated by

$$\sum_{t=0}^{T-1}[\nabla_\theta \log \pi(\boldsymbol{a}_t|\boldsymbol{o}_t;\theta)\big(r_t + \gamma V(\boldsymbol{o}_{t+1};\theta_v) - V(\boldsymbol{o}_t;\theta_v)\big)], \quad (20)$$

where $r_t + \gamma V(\boldsymbol{o}_{t+1};\theta_v) - V(\boldsymbol{o}_t;\theta_v)$ is an estimate of the advantage function. To further enhance the sample efficiency, the multi-steps estimate is utilized as follows

$$A(\boldsymbol{o}_t, \boldsymbol{a}_t;\theta_v) = \sum_{\tau=t}^{t+k-1} \gamma^{\tau-t} r_\tau + \gamma^k V(\boldsymbol{o}_{t+k};\theta_v) - V(\boldsymbol{o}_t;\theta_v), \quad (21)$$

where k is upper-bounded by the allowed maximal steps in one update t_{max}. The policy $\pi(\cdot \mid \cdot;\theta)$ and value function $V(\cdot;\theta_v)$ are normally called *actor* and *critic*, parameters of which are eventually updated every t_{max} actions with following gradients [25]

$$\begin{aligned} d\theta &= \sum_{\tau=t}^{t+t_{max}} \nabla_\theta \log \pi(\boldsymbol{a}_\tau|\boldsymbol{o}_\tau;\theta)\, A(\boldsymbol{o}_\tau, \boldsymbol{a}_\tau;\theta_v) \\ d\theta_v &= \sum_{\tau=t}^{t+t_{max}} \partial A(\boldsymbol{o}_\tau, \boldsymbol{a}_\tau;\theta_v)^2/\partial\theta_v \end{aligned} \quad (22)$$

C. Vectorized Advantage Actor-Critic

The benchmark A2C algorithm regards the action $\boldsymbol{a}_t$ as a whole, considering only the total reward r_t and overlooking the contribution of each element in action $\boldsymbol{a}_t$. This would introduce a large variance of the estimate of advantage and substantially impede the exploration for optimal action, especially faced with problems with high-dimensional action space. For example, under our proposed scheme, u_t^{DG} is presumed to be a bad control signal for controlling DGs, while other elements of action $\boldsymbol{a}_t$ are good enough such that the overall reward r_t is better than expected, i.e., the advantage A is large than zero. In this case, despite this action is terrible, the *actor* network would be updated to increase the probability of this action under the same observation, because the advantage A includes effects from other elements of action $\boldsymbol{a}_t$.

To tackle this issue, we introduce the credit assignment mechanism to decompose the total reward r_t [32]. According to the definition of r_t in (11), some components of total reward are attributable, which provides the probability to assign the total reward to each element in action. Specifically, the generation cost of DGs C_t^{DG} is only determined by the control signal of DGs, independent to other elements of action $\boldsymbol{a}_t$. Similarly, the user dissatisfaction $C_t^{EV_i}$ and thermal comfort loss $C_t^{AC_j}$ are only influenced by the charging power $P_t^{EV_i}$ and AC input power $P_t^{AC_j}$, respectively. In contrast, according to power balance constraint (6), the degradation cost of BESs C_t^{BES} is affected by all elements of action $\boldsymbol{a}_t$, which cannot be assigned to one specific element in action. In this sense, we create a shaped reward for each element of action that reflect its own consequences on the total reward by removing a large amount of the noise created by other elements of action

$$\begin{cases} r_t^{DG} = -C_t^{DG} - C_t^{BES}, \\ r_t^{EV_i} = -C_t^{EV_i} - C_t^{BES}, i = 1, \dots, M, \\ r_t^{AC_j} = -C_t^{AC_j} - C_t^{BES}, j = 1, \dots, N. \end{cases} \quad (23)$$

This credit assignment method is known as difference rewards [33]. Furthermore, the shaped rewards for all elements of action are formed as a vector

$$\boldsymbol{r}_t = \left[r_t^{DG}, r_t^{EV_1}, \dots, r_t^{EV_M}, r_t^{AC_1}, \dots, r_t^{AC_N}\right]. \quad (24)$$

Accordingly, the discounted cumulative reward is modified as a vector $\boldsymbol{R}_t = \sum_{\tau=t}^{T-1} \gamma^{\tau-t} \boldsymbol{r}_\tau$. The policy is extended as $\boldsymbol{\pi}(\cdot \mid \cdot;\theta): \mathcal{A} \times \Omega \to \mathbb{R}^K$, which manifests the independent probability distribution of each element in action $\boldsymbol{a}_t$. The value function is augmented as vector function $\boldsymbol{V^\pi}(\cdot): \Omega \to \mathbb{R}^K$ to evaluate $\boldsymbol{V^\pi}(\boldsymbol{o}_t) = \mathbb{E}_{\boldsymbol{a}_{\tau \geq t} \sim \boldsymbol{\pi}}[\boldsymbol{R}_t|\boldsymbol{o}_t]$. On this basis, the multi-step advantage is redefined as follows,

$$\boldsymbol{A}(\boldsymbol{o}_t, \boldsymbol{a}_t;\theta_v) = \sum_{\tau=t}^{t+k-1} \gamma^{\tau-t} \boldsymbol{r}_\tau + \gamma^k \boldsymbol{V}(\boldsymbol{o}_{t+k};\theta_v) - \boldsymbol{V}(\boldsymbol{o}_t;\theta_v), \quad (25)$$

which reflects the advantages of all elements in the action $\boldsymbol{a}_t$, rather than the joint advantage of action $\boldsymbol{a}_t$.

To accommodate the vector operations in parameter updates, the product of scalars in the first line of (22) is replaced by the inner product of vectors; the square of scalar in the second line of (22) is replaced by the square of the norm of vector. The detail of VA2C is expressed as Algorithm 1.

Algorithm 1 VA2C algorithm

Randomly initialize parameter vectors θ and θ_v
$t \leftarrow 0$, $T_{total} \leftarrow 0$
repeat
  Reset gradients: $d\theta \leftarrow 0$ and $d\theta_v \leftarrow 0$
  for each environment thread **do**
    Receive observation $\boldsymbol{o}_t$, $t_{start} \leftarrow t$
    repeat
      Sample $\boldsymbol{a}_t$ according to distribution $\boldsymbol{\pi}(\boldsymbol{o}_t;\theta)$
      Execute $\boldsymbol{a}_t$ and receive $\boldsymbol{o}_{t+1}$ and r_t
      Perform credit assignment and get vector $\boldsymbol{r}_t$
    until $t = T$ **or** $t - t_{start} == t_{max}$
      $\boldsymbol{R} = \begin{cases} \boldsymbol{0} & \text{for } t = T \\ \boldsymbol{V}(\boldsymbol{o}_t;\theta_v) & \text{for } t \neq T \end{cases}$
    for $\tau \in \{t-1, \dots, t_{start}\}$ **do**
      $\boldsymbol{R} \leftarrow \boldsymbol{r}_\tau + \gamma \boldsymbol{R}$
      Accumulate gradients with respect to θ:
      $d\theta \leftarrow d\theta + \nabla_\theta\left[\log \boldsymbol{\pi}(\boldsymbol{a}_\tau|\boldsymbol{o}_\tau;\theta) \cdot \left(\boldsymbol{R} - \boldsymbol{V}(\boldsymbol{o}_\tau;\theta_v)\right)\right]$
      Accumulate gradients with respect to θ_v:
      $d\theta_v \leftarrow d\theta_v + \partial\|\boldsymbol{R} - \boldsymbol{V}(\boldsymbol{o}_\tau;\theta_v)\|^2/\partial\theta_v$
    end for
  end for
  Perform update of θ using $d\theta$ and of θ_v using $d\theta_v$
until $T_{total} > T_{max}$

It is notable that the developed VA2C algorithm is a general DRL algorithm that can be applied to handle problems with high-dimensional continuous action space and (partially) decomposable reward.

D. VA2C for Privacy Preserving Load Control

As mentioned at the beginning of this section, the partial observability brings challenge for policy-based DRL algorithms to addressing the formulated POMDP problem. To tackle this issue, inspired by the advances of RNN in sequence tasks, we employ the RNN architecture to enhance the representation capability of whole neural networks.

In this work, as shown in Fig.2, the gated recurrent unit (GRU) RNN is concatenated with *actor* and *critic* network. Similar to long short-term memory (LSTM), GRU has been proved to manage to the vanishing gradient problem [34]. Moreover, GRU includes two gates, maintaining simpler structure compared with LSTM consisting of three gates, while comparable to LSTM on performance [35]. Parameterized by θ_r, the GRU is mathematically expressed as follows,

$$\boldsymbol{h}_t = GRU(\boldsymbol{h}_{t-1}, \boldsymbol{o}_t; \theta_r). \tag{26}$$

where $\boldsymbol{h}_{t-1}$, the hidden state at last time slot, together with current observation $\boldsymbol{o}_t$, serves as the input of GRU. With the benefit of the recurrent structure, GRU can learn to memory the important information at previous time slots.

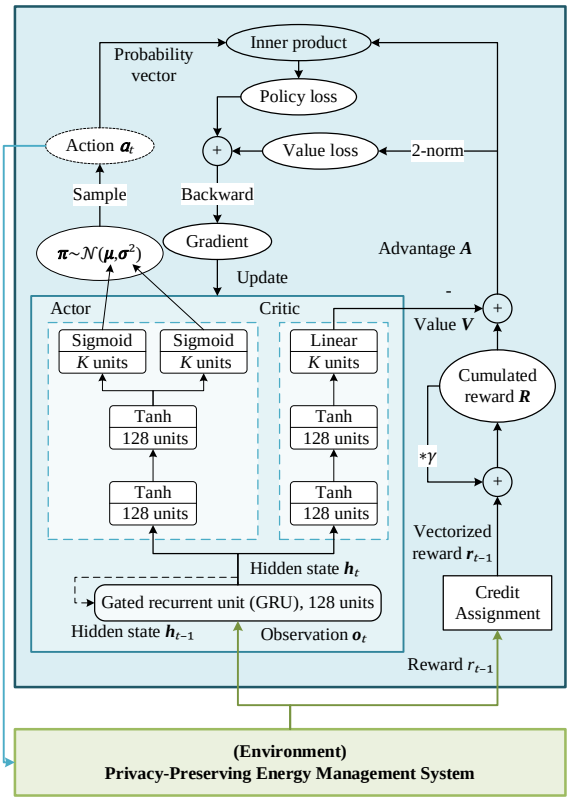

Fig. 2. Work flow of VA2C algorithm for privacy preserving load control scheme.

In practice, the policy is specified as the multivariate Gaussian distribution with diagonal covariance matrix $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, where $\boldsymbol{\mu} = [\mu_1, \dots, \mu_K]$ and $\boldsymbol{\sigma}^2 = [\sigma_1^2, \dots, \sigma_K^2]$ are K-dimensional vectors, determining the mean and variance of probability distribution for each element of action. As shown in Fig. 2, we use three fully connected layers that have one sigmoid output for mean $\boldsymbol{\mu}$ and another sigmoid output for variance $\boldsymbol{\sigma}^2$, with all non-output layers shared. Similarly, we use three fully connected layers that have one linear output for value function, with one GRU layer shared with policy network.

The work flow of the developed VA2C algorithm for privacy preserving load control scheme is depicted in Fig. 2. The current observation $\boldsymbol{o}_t$, along with the hidden state at last time slot $\boldsymbol{h}_{t-1}$, serves as input of GRU layer. Then, GRU layer outputs current hidden state $\boldsymbol{h}_t$ which is fed into the *actor* and *critic* network to generate policy $\boldsymbol{\pi}$ and value $\boldsymbol{V}$. The action $\boldsymbol{a}_t$, sampled according to the distribution $\boldsymbol{\pi}$ is performed in the environment. Meanwhile, the reward r_{t-1} is expanded into vector $\boldsymbol{r}_{t-1}$ after passing through credit assignment module, and provided to calculate the discounted cumulative reward $\boldsymbol{R}$. Next, advantage $\boldsymbol{A}$, the difference between $\boldsymbol{V}$ and $\boldsymbol{R}$ is processed to calculate value loss (the 2-norm of advantage) and policy loss (the inner product of advantage and probability vector). By backward propagation, the gradients of total loss are obtained, and then used to update the whole neural network.

IV. Performance Evaluation

In this section, the performance of the developed VA2C algorithm is verified, and the effectiveness of the proposed privacy preserving load control scheme is evaluated. First, the simulation experiment setup is described. Then, the baseline schemes used for performance comparisons are introduced. Finally, the simulation results of comparisons between different schemes and corresponding discussions are provided.

A. Experiment Setup and Implementation

1) Environment: We consider the load control during one day with 5 minutes of duration of one time slot, such that the time horizon T is 288. Put differently, the number of transitions in one episode is 288. The real-world power data and temperature data are used to describe power of PVs, power basic loads and outdoor temperature, which are obtained from Pecan Street Database [28] and NOAA [29]. Based on the analysis of the real EV charging power data and AC power data [28], the arrival and departure time of each EV and the start and end usage time of each AC are satisfied with distinct uniform distributions.

For simulation, the transition functions and cost functions are specified in Appendix. Note that the specified functions would not be served as prior knowledge by microgrid operator. Therefore, the problems with other functions or even without explicit transition dynamics can also be well addressed by our model-free algorithm.

2) Algorithm: We employ 8 threads to interact with the environment. To prevent over-fitting, the 100 days of historical data are randomly divided into training set, validation set and test set, the proportions of which are 80%, 10% and 10%, respectively. Considering the total time steps of 288, the discount factor γ is set to be 0.95, which can avoid learning a myopic policy while reducing the variance of estimate during

learning. According to empirical results [37], the learning rates of *actor* network and *critic* network parameter are set to be 0.0002 and 0.0001. The number of steps per update t_{max} is set to be 24. The DRL algorithms are implemented using PyTorch in Python. The case studies are carried out on a server with an 8-core AMD Ryzen 7 3700X processor and one single GeForce RTX 2080 GPU.

B. Baselines

The performance of proposed scheme is compared with four baseline schemes as follows.

Baseline1: The privacy preserving scheme with benchmark DRL algorithm, i.e., A2C. The network of A2C is identical in structure to the network portrayed in Fig.2, except the number of units in output layer of *critic* network. We intend to manifest the superiority of the developed VA2C algorithm through the comparison with *Baseline1*.

Baseline2: The scheme regardless of privacy issues. Since the state of microgrid is fully observed by the microgrid operator under this scheme, we implement this scheme with VA2C algorithm without recurrent layer. The cost incurred by privacy preservation can be revealed through the comparison with *Baseline2*.

Baseline3: The existing privacy preserving scheme. A typical AC control strategy without the information about indoor temperature, set-point temperature control [17], is conducted. Under this scheme, the microgrid operator decides the set-point temperature for each AC, and then the power of each AC is set by HEMS according to the set-point temperature: the AC operates at maximum power when the indoor temperature is higher than set-point temperature; turns off otherwise. This strategy can be mathematically represented as follows

$$P_t^{AC_j} = \begin{cases} P_{max}^{AC_j}, T_t^{AC_j} > T_{t,set}^{AC_j} \\ 0, \quad T_t^{AC_j} \le T_{t,set}^{AC_j}, \end{cases} \tag{27}$$

where $T_{t,set}^{AC_j}$ denotes the set-point temperature of AC_j. We implement this scheme with the developed VA2C algorithm, intending to fairly compare this AC control strategy with the AC direct power control.

Baseline4: The privacy preserving scheme with the aim of maximizing the user comfort. Under this scheme, the microgrid operator determines the power of EVs and ACs without the consideration of operation cost. We implement this scheme by conducting following credit assignment,

$$\begin{cases} r_t^{DG} = -C_t^{DG} - C_t^{BES} \\ r_t^{EV_i} = -C_t^{EV_i}, i = 1, \dots, M \\ r_t^{AC_j} = -C_t^{AC_j}, j = 1, \dots, N \end{cases} \tag{28}$$

In comparison with (23), the rewards assigned to actions that control EVs and ACs in (28) dismiss the operation cost including generation cost of DGs and degradation cost of BESs, such that the learned policy aims to minimize the user comfort loss.

C. Results

To evaluate the performances of different schemes during the learning process, in Fig.3, we depict the mean episode reward ($R = \sum_{t=0}^{T-1} r_t$) on the validation set with 10 different random seeds. After learning process, we evaluate the trained policies over the test set, and the results are shown in Table I. The operation cost includes the generation cost of DGs and the degradation cost of BESs, while the user comfort loss consists of the user dissatisfaction caused by EV charging scheduling and the thermal comfort loss caused by ACs.

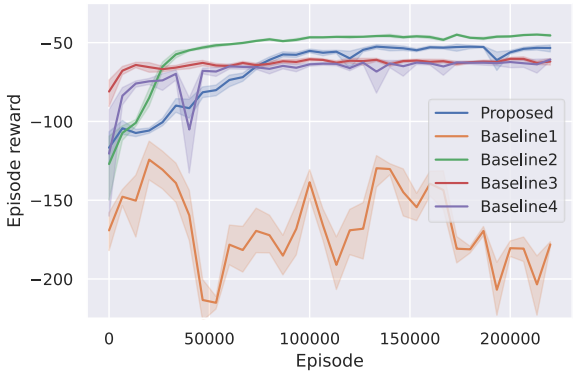

Fig. 3. Performances of different schemes during learning. The shaded region represents 90% confidence interval.

TABLE I
AVERAGE DAILY COST OVER TEST SET

Scheme	Operation cost	User comfort loss	Total cost
Proposed	68.03	20.59	88.62
Baseline1	54.26	122.81	177.07
Baseline2	57.23	18.36	75.61
Baseline3	87.12	30.14	117.26
Baseline4	115.37	5.12	120.49

1) Algorithmic Superiority: We first focus on the comparison between the developed VA2C algorithm and the benchmark A2C algorithm. Note that *Baseline1* is implemented with A2C while other schemes are implemented with proposed VA2C. It can be observed from Fig.3 that the performance of all schemes except *Baseline1* converge, showing the benefits of VA2C in the stability of learning.

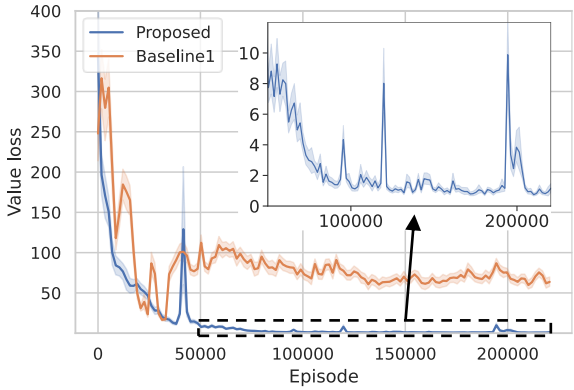

Fig. 4. Comparison of value loss during learning between the developed VA2C algorithm in proposed scheme and benchmark A2C algorithm in *Baseline1*.

The value estimate loss comparison between VA2C and A2C is demonstrated in Fig. 4. The mean value estimate loss of A2C is larger than 50 during learning process, revealing the large bias of value estimates, which leads to unstable learning shown in Fig. 3. In contrast, the mean value estimate loss of VA2C decreases rapidly and eventually converges to below 2.

Furthermore, Table I shows that the average daily cost of proposed scheme implemented by VA2C is reduced by 50% than that of *Baseline1* implemented by A2C. Especially, the user comfort loss is reduced by 83%. These results demonstrate that the developed VA2C algorithm significantly outperforms the benchmark A2C algorithm on the privacy preserving load control problem.

The superior performance of VA2C could be explained from two perspectives. First, with the benefit of credit assignment, the value function is designed to estimate a series of rewards rather than a total reward, which reduces the variance of estimate. Second, the policy network is updated in the direction of maximizing the advantages associated with the elements of action rather than the advantage of the joint action, which improves the efficiency and stability of training.

2) Cost Incurred by Privacy Preservation: In this part, we focus on the performance comparison with existing privacy preserving scheme (*Baseline3*) and scheme regardless of privacy preservation (*Baseline2*). The performances during learning process are shown in Fig. 3. After convergence, the performance of proposed scheme is better than *Baseline3* and slightly worse than *Baseline2*, which is also reflected in Table I. Using *Baseline2* as a benchmark, the total cost of proposed scheme increases by 14.7%, compared to a 55.1% increase in that of *Baseline3*. Especially, Fig. 5 demonstrates the overall AC control performances: the performances of proposed privacy preserving scheme are relatively close to that of the scheme regardless of privacy preservation; compared with another privacy preserving scheme *Baseline3*, ACs under the proposed scheme cause the lower thermal comfort loss while only consume 60% of the energy consumed by ACs under *Baseline3*.

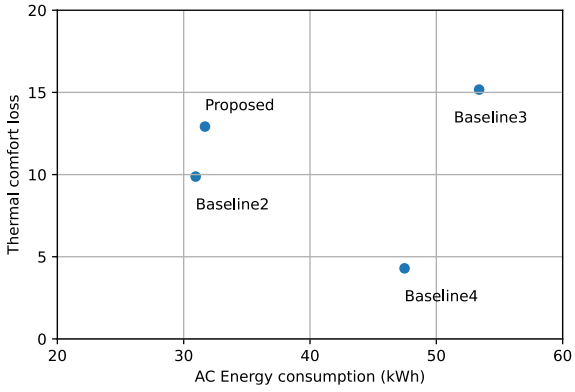

Fig. 5. Thermal comfort loss and energy consumption of ACs under different scheme except *Baseline1*.

To specify, Fig. 6 shows the indoor temperature curves on a typical day selected from test set. On the one hand, compared with *Baseline2*, the proposed scheme takes a risk of over-control for AC, due to the absence of the indoor temperature information for decision-making. This may cause user comfort loss and increase energy consumption. One the other hand, compared with another privacy preserving scheme, the AC control under proposed scheme is more robust and energy efficient, primarily because there are only two mode for the power of ACs, i.e., zero and maximum, under the set-point temperature control of *Baseline3*.

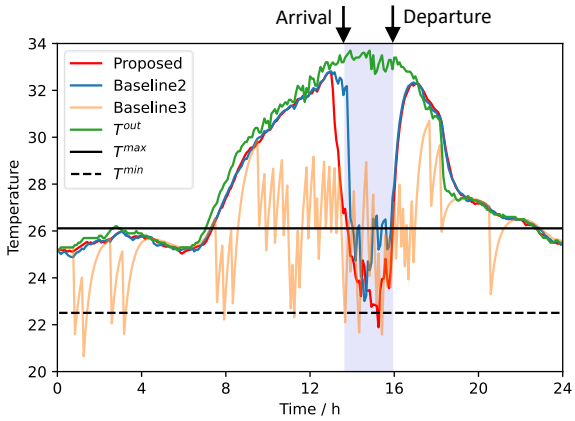

Fig. 6. Indoor temperature curves under proposed scheme, *Baseline2* and *Baseline3*.

We notice an interesting phenomenon where the learned control strategy reduces the indoor temperature to zero comfort loss zone in advance without knowing the accurate arrival time of the occupant. This advance primarily benefits from the RNN architecture which possesses the capability to remember the previous observation. Overall, compared to *Baseline2* and *Baseline3*, the proposed scheme protects user privacy at a relatively small cost.

3) Flexibility and Scalability: The goal of the proposed scheme is to minimize the total cost including operation cost of the microgrid and user comfort loss, while preference can be considered and implemented by designing appropriate credit assignment in the developed VA2C algorithm. For example, *Baseline4* aims to minimize user comfort loss by employing specific credit assignment shown as (28). The results in Table I show that *Baseline4* minimizes the comfort loss, even 72% less than that of the scheme regardless of privacy preservation (*Baseline2*). Meanwhile, the operation cost of microgrid under *Baseline4* is higher than other schemes.

To further compare the control effects of the proposed scheme and *Baseline4*, the total charging power of EVs and total working power of ACs during a typical day are presented in Fig. 7. It can be observed that the EV charging scheduled by the proposed scheme matches well with the output of PVs, consequently saving the operation cost of DGs and BESs. In contrast, the EVs scheduled by *Baseline4* aims to charge up as quickly as possible and prevent the user dissatisfactions. The control for ACs also reflects the goal of *Baseline4*. Compared to the proposed scheme, the ACs controlled by *Baseline4* cause the 29% of thermal comfort loss with a 50% increase in energy consumption.

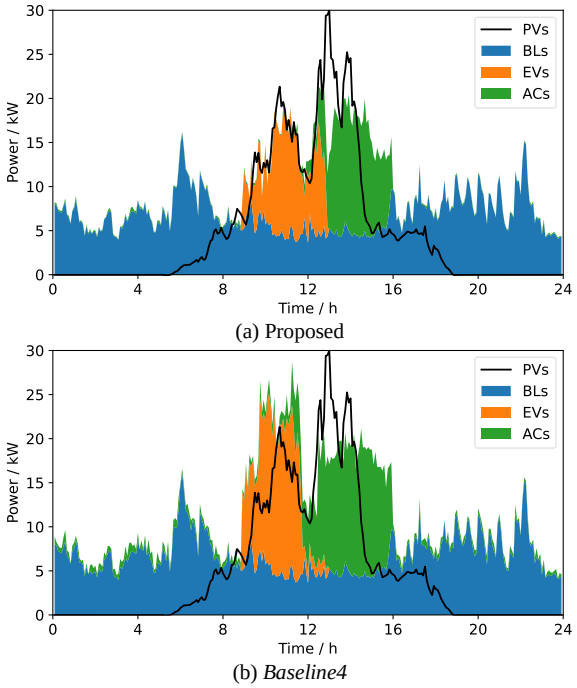

Fig. 7. Power of PVs and loads under proposed scheme and *Baseline4*.

To specify the EV charging scheduling under the proposed scheme and *Baseline4*, the dynamics of charging demands for two EVs during a typical day are presented in Fig. 8. It can be observed that under *Baseline4* both EV_3 and EV_4 are charged at almost maximum power until the charging demands are satisfied. This charging scheduling strategy aims to adequately satisfy the users' charging demand regardless of the power balance and operation cost of the microgrid. In contrast, under the proposed scheme, two EVs are charged at modest and variable power, thereby leveraging the generation of PVs and reducing the operation cost. However, this charging scheduling strategy has the potential to causing user's dissatisfaction. For example, EV_4 is not well charged at departure time under the proposed scheme.

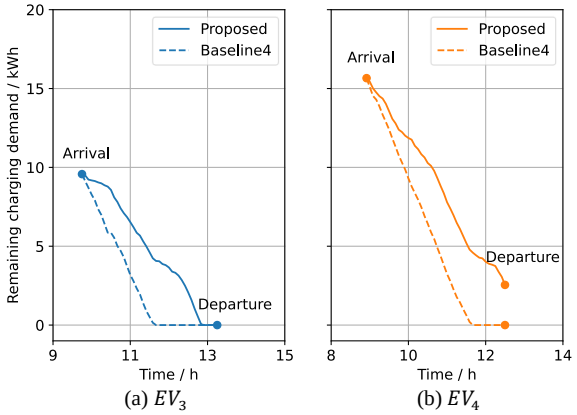

Fig. 8. The charging scheduling for EV_3 and EV_4 under the proposed scheme and *Baseline4*.

The comparison between *Baseline4* and the proposed scheme demonstrates that diverse control effects can be implemented by designing proper credit assignment.

V. Conclusion and Future Work

In this paper, the privacy preserving load control problem is investigated. We propose a two-layer interactive architecture that achieves effective control while preserving the user privacy data through the interaction of microgrid operator and HEMS. We formulate this issue as a POMDP problem and develop the VA2C algorithm to address the challenge caused by the high-dimensional continuous action space and integrate RNN to cope with the partial observability due to privacy issues. Through numerical simulation, we demonstrate that the developed VA2C algorithm can improve the stability of learning and the control performance. Moreover, the proposed scheme achieves privacy protection with a 14.7% increase of total cost, compared to a 55.1% increase of the existing privacy preserving method. Particularly, the proposed scheme enables pre-cooling without the information about arrival time of occupants, outperforming the scheme with full state in this respect. Finally, we demonstrate that the developed algorithm can be flexibly adjusted to achieve diverse control effect, such as lowest user comfort loss.

The work of this paper provides a novel idea for privacy protection in load control, that is, using DRL to achieve effective control without private information. However, some users' privacy information is still required by microgrid operator, such as current usage status of ACs and current charging demands of EVs. In our future work, one the one hand, we will explore the tradeoff between privacy preservation and control effect by designing the privacy as optimization objectives or decision variables; on the other hand, the multi-agent DRL algorithms will be applied to the load control problems, intending to fundamentally eliminate the privacy issues. Moreover, in many other scenarios, e.g., centralized power trading, peer to peer trading, etc., this idea can also be applied to achieve optimal decision-making while effectively preserving privacy. In our future work, the application method of this privacy preserving idea in more scenarios will be further investigated.

Appendix

The transition functions and cost functions used in the simulation are specified as follows.

The transition function of power of DGs $f^{DG}(\cdot,\cdot)$ is typically defined as [2]

$$f^{DG}(P,u) = \left(1 - \frac{\Delta t}{T^{DG}}\right) P + \frac{P_{max}^{DG} \Delta t}{T^{DG}} u, \tag{29}$$

where T^{DG} and P_{max}^{DG} are the time constant and maximum power of DGs, respectively. The transition function of SOC is specified as [8]

$$f^{BES}(SOC,P) = \begin{cases} SOC + \frac{P\Delta t}{Q_s}\eta_{in}, P \geq 0, \\ SOC + \frac{P\Delta t}{Q_s \eta_{out}}, P < 0, \end{cases} \tag{30}$$

where Q_s, η_{in} and η_{out} are the capacity, charging and discharging efficiency coefficients of BESs. The transition function of energy demand during EV charging is unified as a linear model [21] $f^{EV_i}(E,P) = E - P\Delta t, i = 1, \ldots, M$. The transition function of indoor temperature is represented by the

equivalent thermal parameter model [7] as follows,

$$f^{AC_j}(T, T^{out}, P) = T + \left(1 - e^{-a_j^{AC}\Delta t}\right)\left(T^{out} - T - b_j^{AC}P\right), \quad (31)$$

where a_j^{AC} and b_j^{AC} are the coefficients determined by the thermal characteristics of corresponding room and AC, such as room area, heat preservation performance and cooling efficiency.

For simplicity, $g^{DG}(\cdot)$ is defined as quadratic form [2] $g^{DG}(P) = \lambda_1^{DG}P + \lambda_2^{DG}P^2$, where λ_1^{DG} and λ_2^{DG} are coefficients of DGs.

We formulate the degradation cost function of BESs $g^{BES}(\cdot,\cdot)$ as a rigorous integral model

$$g^{BES}(SOC, P) = Q_s \left| \int_{SOC}^{f^{BES}(SOC,P)} h^{BES}(x)dx \right|, \quad (32)$$

where $h^{BES}(\cdot)$ denotes marginal degradation cost function per kWh associated with SOC, and are specified as piecewise linear function [36]

$$h^{BES}(SOC) = \lambda_1^{BES} + \begin{cases} (\lambda_1^{BES} - \lambda_2^{BES})(1 - 2SOC), & \text{if} \quad 0.5 \le SOC \le 1, \\ 0, & \text{if} \quad 0 \le SOC < 0.5, \end{cases} \quad (33)$$

where λ_1^{BES} and λ_2^{BES} are maximum and minimum degradation cost per kWh, corresponding to SOC of 0 and SOC of 1, respectively. The specific values of λ_1^{BES} and λ_2^{BES} are also referred to [36].

The user dissatisfaction function of EV charging is also defined as quadratic form $g^{EV_i}(E) = \lambda_{i,1}^{EV}E + \lambda_{i,2}^{EV}E^2$, where $\lambda_{i,1}^{EV}$ and $\lambda_{i,2}^{EV}$ are coefficients associated with the charging of EV_i. The comfort loss function $g^{AC_j}(\cdot)$ is defined as

$$g^{AC_j}(T) = \begin{cases} 0, T_{min}^j \le T \le T_{max}^j, \\ \lambda_{j,1}^{AC} \exp\left[\lambda_{j,1}^{AC}\left(T - T_{max}^j\right)\right], T > T_{max}^j, \\ \lambda_{j,1}^{AC} \exp\left[\lambda_{j,2}^{AC}\left(T_{min}^j - T\right)\right], T < T_{min}^j, \end{cases} \quad (34)$$

where $[T_{min}^j, T_{max}^j]$ is the zero-loss temperature zone, μ_j^{AC} and λ_j^{AC} are weight coefficients associated with the comfort loss caused by AC_j.

The important environment parameters are configured in Table II, and other parameters are provided in Table III, Table IV and Table V.

TABLE II
Environment parameter settings

Parameter	Value	Parameter	Value	Parameter	Value
M	5	N	10	η_{in}	0.95
η_{out}	0.95	T^{DG}	20 min	P_{max}^{DG}	20 kW
λ_1^{DG}	0.5	λ_2^{DG}	0.0125	Q_s	50 kWh
λ_1^{BES}	0.13	λ_2^{BES}	0.05		

TABLE III
PARAMETERS AND DISTRIBUTIONS ASSOCIATED WITH THE DYNAMICS OF INDOOR TEMPERATURE

j	$P_{max}^{AC_j}$	a_j^{AC}	b_j^{AC}	$t_e^{AC_j}$	$t_l^{AC_j}$
1	1 kW	2.50	17.7	$\mathcal{U}$(11:00,12:00)	$\mathcal{U}$(13:00,15:00)
2	1 kW	2.27	15.4	$\mathcal{U}$(12:45,13:45)	$\mathcal{U}$(14:45,16:45)
3	2 kW	2.91	8.50	$\mathcal{U}$(11:05,12:05)	$\mathcal{U}$(13:05,15:05)
4	2 kW	2.38	8.45	$\mathcal{U}$(11:55,12:55)	$\mathcal{U}$(13:55,15:55)
5	2 kW	2.85	8.77	$\mathcal{U}$(12:50,13:50)	$\mathcal{U}$(14:50,16:50)
6	2 kW	1.71	8.25	$\mathcal{U}$(12:20,13:20)	$\mathcal{U}$(14:20,16:20)
7	2 kW	1.71	8.50	$\mathcal{U}$(10:35,11:35)	$\mathcal{U}$(12:35,14:35)
8	3 kW	2.71	5.13	$\mathcal{U}$(12:25,13:25)	$\mathcal{U}$(14:25,16:25)
9	3 kW	2.10	5.03	$\mathcal{U}$(12:40,13:40)	$\mathcal{U}$(14:40,16:40)
10	3 kW	1.75	5.86	$\mathcal{U}$(12:50,13:50)	$\mathcal{U}$(14:50,16:50)

TABLE IV
PARAMETERS OF THERMAL LOSS FUNCTION OF ACS

j	T_j^{min}	T_j^{max}	$\lambda_{j,2}^{AC}$	$\lambda_{j,1}^{AC}$
1	23.1	26.3	0.396	0.352
2	21.0	24.7	0.349	0.694
3	24.4	26.7	0.317	0.467
4	22.8	26.3	0.289	0.775
5	22.5	26.1	0.331	0.454
6	22.8	26.7	0.376	0.421
7	22.6	26.6	0.212	1.24
8	20.8	24.5	0.215	1.10
9	23.9	26.5	0.263	1.03
10	20.6	24.0	0.303	0.670

TABLE V
PARAMETERS AND DISTRIBUTIONS ASSOCIATED WITH EV CHARGING

i	$P_{max}^{EV_i}$	$\lambda_1^{EV_i}$	$\lambda_2^{EV_i}$	$t_a^{EV_i}$	$t_d^{EV_i}$
1	3.4 kW	1.12	0.020	$\mathcal{U}$(8:15,9:15)	$\mathcal{U}$(11:15,13:15)
2	3.4 kW	1.70	0.034	$\mathcal{U}$(8:55,9:55)	$\mathcal{U}$(11:55,13:55)
3	6.8 kW	1.57	0.023	$\mathcal{U}$(8:00,9:00)	$\mathcal{U}$(11:00,13:00)
4	6.8 kW	1.56	0.033	$\mathcal{U}$(9:20,10:20)	$\mathcal{U}$(12:20,14:20)
5	2.0 kW	1.14	0.029	$\mathcal{U}$(9:10,10:10)	$\mathcal{U}$(12:20,14:20)

REFERENCES

[1] O. Erdinç, A. Taşcikaraoğlu, N. G. Paterakis, and J. P. S. Catalão, "Novel incentive mechanism for end-users enrolled in DLC-based demand response programs within stochastic planning context," *IEEE Trans. Ind. Electron.*, vol. 66, no. 2, pp. 1476-1487, Feb. 2019.

[2] H. Hua, Y. Qin, C. Hao, and J. Cao, "Stochastic optimal control for energy internet: a bottom-up energy management approach," *IEEE Trans. Ind. Inform.*, vol. 15, no. 3, pp. 1788-1797, Mar. 2019.

[3] K. Paridari, A. Parisio, H. Sandberg, and K. H. Johansson, "Robust scheduling of smart appliances in active apartments with user behavior uncertainty," *IEEE Trans. Autom. Sci. Eng.*, vol. 13, no. 1, pp. 247-259, Jan. 2016.

[4] T. Morstyn, N. Farrell, S. J. Darby, and M. D. McCulloch, "Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants," *Nat. Energy*, vol. 3, no. 2, pp. 94-101, Feb. 2018.

[5] M. Vanouni and N. Lu, "Improving the centralized control of thermostatically controlled appliances by obtaining the right information," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 946-948, Mar. 2015.

[6] Q. Cui, X. Wang, X. Wang, and Y. Zhang, "Residential appliances direct load control in real-time using cooperative game," *IEEE Trans. Power Syst.*, vol. 31, no. 1, pp. 226-233, Jan. 2016.

[7] F. Luo, W. Kong, G. Ranzi, and Z. Y. Dong, "Optimal home energy management system with demand charge tariff and appliance operational dependencies," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 4-14, Jan. 2020.

[8] L. Yu *et al.*, "Deep reinforcement learning for smart home energy management," *IEEE Internet of Things J.*, vol. 7, no. 4, pp. 2751-2762, Apr. 2020.

[9] Y. Du, H. Zandi, O. Kotevska, K. Kurte, J. Munk, K. Amasyali, E. Mckee, and F. Li, "Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning," *Appl. Energy*, vol. 281, 2021.

[10] Z. Liu, Q. Wu, S. Huang, L. Wang, M. Shahidehpour and Y. Xue, "Optimal day-ahead charging scheduling of electric vehicles through an aggregative game model," *IEEE Trans. on Smart Grid*, vol. 9, no. 5, pp. 5173-5184, Sept. 2018.

[11] H. Hou, *et al.*, "Multi-objective economic dispatch of a microgrid considering electric vehicle and transferable load," *Appl. Energy*, vol. 262, 2020.
[12] J. Mohammadi, G. Hug, and S. Kar, "A fully distributed cooperative charging approach for plug-in electric vehicles," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3507-3518, Jul. 2018.
[13] Y. Yang, Q. Jia, G. Deconinck, X. Guan, Z. Qiu, and Z. Hu, "Distributed coordination of EV charging with renewable energy in a microgrid of buildings," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6253-6264, Nov. 2018.
[14] M. Ahmadi, J. M. Rosenberger, W. J. Lee, and A. Kulvanitchaiyanunt, "Optimizing load control in a collaborative residential microgrid environment," *IEEE Trans. Smart Grid*, vol. 6, no. 3, pp. 1196-1207, May 2015.
[15] N. Ahmed, M. Levorato, and G. P. Li, "Residential consumer-centric demand side management," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4513-4524, Sept. 2018.
[16] A. Baniasadi, D. Habibi, O. Bass and M. A. S. Masoum, "Optimal real-time residential thermal energy management for peak-load shifting with experimental verification," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5587-5599, Sept. 2019.
[17] A. Halder, X. Geng, P. R. Kumar, and L. Xie, "Architecture and algorithms for privacy preserving thermal inertial load management by a load serving entity," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 3275-3286, Jul. 2017.
[18] Y. Gong, Y. Cai, Y. Guo, and Y. Fang, "A privacy-preserving scheme for incentive-based demand response in the smart grid," *IEEE Trans. Smart Grid*, vol. 7, no. 3, pp. 1304-1313, May 2016.
[19] Q. Zhang, K. Dehghanpour, Z. Wang, and Q. Huang, "A learning-based power management method for networked microgrids under incomplete information," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1193-1204, Mar. 2020.
[20] H. Matthew and P. Stone, "Deep recurrent Q-learning for partially observable MDPs," In *AAAI Fall Symposium on Sequential Decision Making for Intelligent Agents*, Nov. 2015.
[21] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-free real-time EV charging scheduling based on deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5246-5257, Sept. 2019.
[22] B. Wang, Y. Li, W. Ming, and S. Wang, "Deep Reinforcement Learning Method for Demand Response Management of Interruptible Load," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3146-3155, Jul. 2020.
[23] T. P. Lillicrap, *et al.*, "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learning Representations*, San Juan, Puerto Rico, 2016.
[24] L. Lin, X. Guan, Y. Peng, N. Wang, S. Maharjan, and T. Ohtsuki, "Deep reinforcement learning for economic dispatch of virtual power plant in internet of energy," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6288-6301, Jul. 2020.
[25] M. Volodymyr, *et al.*, "Asynchronous methods for deep reinforcement learning," in *proc. Int. Conf. Mach. Learning*, New York, USA, pp. 1928-1937, 2016.
[26] E. Mocanu et al., "On-line building energy optimization using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3698-3708, Jul. 2019.
[27] T. M. Hansen, E. K. P. Chong, S. Suryanarayanan, A. A. Maciejewski, and H. J. Siegel, "A partially observable markov decision process approach to residential home energy management," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 1271-1281, Mar. 2018.
[28] Pecan Street Database. [Online]. Available: http://www.pecanstreet.org/.
[29] NOAA Data. [Online]. Available: https://www.ncdc.noaa.gov/.
[30] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Overcoming exploration in reinforcement learning with demonstrations," in *Proc. IEEE Int. Conf. Robot. Autom.*, Brisbane, Australia, 2018, pp. 6292–6299.
[31] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. Int. Conf. Machine Learning*, Beijing, China, 2014, pp. 387–395.
[32] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *AAAI Conference on Artificial Intelligence*, 2018.
[33] S. Devlin, L. Yliniemi, D. Kudenko, and K. Tumer, "Potential-based difference rewards for multiagent reinforcement learning," in *International Conference on Autonomous Agents and Multi-agent Systems*, 2014.
[34] K. Cho, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Conference on Empirical Methods in Natural Language Processing*, 2014.
[35] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning*, Dec. 2014.
[36] C. Liu, X. Wang, X. Wu, and J. Guo, "Economic scheduling model of microgrid considering the lifetime of batteries," *IET Generation, Transmission & Distribution*, vol. 3, pp. 759-767, Nov. 2017.
[37] M. Andrychowicz, *et al.*, "What matters in on-policy reinforcement learning? a large-scale empirical study." arXiv preprint arXiv:2006.05990, 2020.

Zhaoming Qin was born in Shandong, P. R. China in 1998. He received the B.Sc. degree in automation from Beihang University, Beijing, P. R. China, in 2019.

He is currently pursuing the master's degree with the Department of Automation, Tsinghua University, Beijing, P. R. China. His current research focuses on reinforcement learning, specifically in the context of smart grids.

Di Liu received the Ph.D. degree in electrical engineering from North China Electric Power University, Beijing, China, in 2020.

He is currently a postdoctoral fellow at Tsinghua University. His research interests include demand side management electricity market and energy internet.

Haochen Hua was born in Jiangsu, P. R. China in 1988. He received the B.Sc. degree in mathematics with finance in 2011, and the Ph.D. degree in mathematical sciences in 2016, both from the University of Liverpool, Liverpool, UK. From 2016 to 2020, he was a Postdoctoral Fellow in the Research Institute of Information Technology, Tsinghua University, Beijing, P. R. China.

Since 2020, he has been a Professor in the College of Energy and Electrical Engineering, Hohai University, Nanjing, P. R. China. His current research interests include energy Internet system modeling, control and optimization.

Junwei Cao received his Ph.D. in computer science from the University of Warwick, Coventry, UK, in 2001. He received his bachelor and master degrees in control theories and engineering in 1998 and 1996, respectively, both from Tsinghua University, Beijing, China.

He is currently Professor and Vice Dean of Research Institute of Information Technology, Tsinghua University, Beijing, China. He is also Director of Open Platform and Technology Division, Tsinghua National Laboratory for Information Science and Technology. Prior to joining Tsinghua University in 2006, he was a Research Scientist at MIT LIGO Laboratory and NEC Laboratories Europe for about 5 years. He has published over 200 papers and cited by international scholars for over 18,000 times. He has authored or edited 8 books. His research is focused on distributed computing technologies and energy/power applications. Prof. Cao is a Senior Member of the IEEE Computer Society and a Member of the ACM and CCF.