This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3124465, IEEE Transactions on Smart Grid

Peer-to-Peer Energy Trading and Energy Conversion in Interconnected Multi-Energy Microgrids Using Multi-Agent Deep Reinforcement Learning

Tianyi Chen, *Student Member, IEEE*, Shengrong Bu*, *Member, IEEE*, Xue Liu, *Fellow, IEEE*, Jikun Kang, F. Richard Yu, *Fellow, IEEE*, and Zhu Han, *Fellow, IEEE*

***Abstract*—A key aspect of multi-energy microgrids (MEMGs) is the capability to efficiently convert and store energy in order to reduce the costs and environmental impact. Peer-to-peer (P2P) energy trading is a novel paradigm for decentralised energy market designs. In this paper, we investigate the external P2P energy trading problem and internal energy conversion problem within interconnected residential, commercial and industrial MEMGs. These two problems are complex decision-making problems with enormous high-dimensional data and uncertainty, so a multi-agent deep reinforcement learning approach combining the multi-agent actor-critic algorithm with the twin delayed deep deterministic policy gradient algorithm is proposed. The proposed approach can handle the high-dimensional continuous action space and aligns with the nature of P2P energy trading with multiple MEMGs. Simulation results based on three real-world MG datasets show that the proposed approach significantly reduces each MG's average hourly operation cost. The impact of carbon tax pricing is also considered.**

***Index Terms*—Multi-energy microgrids, P2P energy trading, energy conversion, multi-agent deep reinforcement learning.**

I. Introduction

MICROGRIDS (MG) are used to address the challenges arising from having a high share of distributed energy resources (DERs) within a local region in modern energy systems. At the distribution network level, a multi-energy microgrid (MEMG) consists of DERs, energy coupling technologies, local active loads and energy storage systems (ESSs). The recent energy coupling technologies, such as hydrogen fuel cells (FCs), water electrolyser (WE) and electric heat pumps (HPs), can be integrated by multiple energy carries together to benefit the energy systems economically and environmentally [1]. Multiple MEMGs can be networked further to improve the efficiency and reliability of the distribution network. However, besides the primary challenges posed by the intermittent nature of DERs, there is an additional difficulty in the stability and operational safety for the network of multiple MEMGs because the deployment including the size and type of DERs varies by location. It is also not realistic to directly control or operate those DERs by a central authority since they may belong to different owners. Peer-to-peer (P2P) energy trading has emerged as a novel paradigm for decentralised energy market designs. P2P energy trading allows the end-users or MGs to join the trading without a central authority unit [2] and offers an opportunity to produce and sell energy at the edge of the network. Correctly modelling and quantifying the P2P energy trading as well as understanding the flexibility of MEMGs are complicated tasks. It involves not only temporal, multi-vector interactions on different networks (e.g., electricity, heat and gas) in response to uncertain energy generation and demand, but also includes potential conflicting trading and operating policies of MGs.

The literature on the P2P energy trading can be classified into five techniques based on the approaches adopted: game theory, auction theory, constrained optimisation, blockchain and deep reinforcement learning (DRL). In [3]–[11], game theory is used to address the P2P energy trading problems in the electricity sector. Some of this work [4], [8], [10], [11] considers trading among multiple MGs, while [3], [5]–[7], [9] considers trading between prosumers. In [12]–[16], game-theoretic approaches are used to solve the P2P energy trading in a multi-energy setting. However, only in [16] did the authors model the P2P energy trading among multiple MEMGs. In [6], [17]–[21], auction-theoretic approaches are used to address the P2P trading problems between prosumers in the electricity sector. In [22], the authors proposed an auction mechanism for energy trading in a multi-energy district. In [23]–[27], constrained optimisation is applied to P2P energy trading under different market and system constraints in the electricity sector. In [28]–[30], blockchain is used to enable secured and decentralised energy trading in the electricity sector. The game-theoretic and auction-theoretic models are mainly solved by traditional constrained optimisation methods such as mixed-integer linear programming (MILP) [31] and alternating direction method of multipliers (ADMM) [32]. Those methods are useful for a lot of complex tasks considering multiple factors and constraints. However, the MILP method assumes linear relationships among factors [31], while ADMM assumes that the problems are regularised and convex [32], which are unrealistic in many cases.

This work was supported by start-up funds provided by Brock University, University of Glasgow Principal's Early Career Mobility fund, NSF CNS-2128368, CNS-2107216, Toyota and Amazon.

T. Chen is with the James Watt School of Engineering, University of Glasgow, United Kingdom (e-mail: t.chen.1@research.gla.ac.uk).

S. Bu is with the Department of Engineering, Brock University, Canada.

X. Liu and J. Kang are with School of Computer Science, McGill University, Canada.

F.R. Yu is with School of Information Technology and Department of Systems and Computer Engineering, Carleton University, Canada.

Z. Han is with the Department of Electrical and Computer Engineering in the University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul, South Korea, 446-701.

* Corresponding author: sbu@brocku.ca

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3124465, IEEE Transactions on Smart Grid

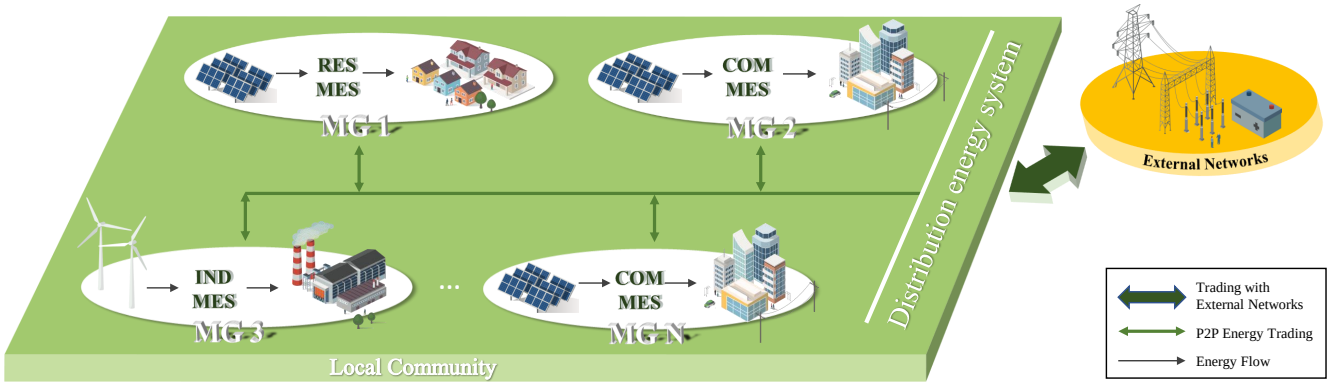


Fig. 1. The framework of P2P energy trading among multiple Multi-energy MGs. RES, COM and IND stand for residential, commercial and industrial.

DRL, combined with deep neural networks (DNNs) and reinforcement learning (RL) techniques, can be powerful tools for addressing the P2P energy trading issues in the network of multiple MEMGs using the trial-and-error mechanism without any extensive feature engineering. In [33], [34], deep Q-learning is used in their corresponding electricity trading problems. In [35], a convolutional neural network (CNN) is used to predict the MG utility while helping the Q-learning algorithm choose the optimal policy for the MG to trade electricity. Deep Q-learning has two major pitfalls: it cannot do well when the environment has a colossal number of actions in continuous action space [36]; it tends to overestimate the Q-value [37].

Previous work on P2P energy trading mainly focuses on the electricity sector. Some work has been done in a multi-energy setting but does not consider both external P2P energy trading and internal energy conversion process. The literature also lacks modelling of different types of MEMGs participating in the P2P energy trading. Considering the varieties of MGs such as residential, commercial and industrial is essential for P2P energy trading and energy conversion within a local community since the energy generation patterns and energy coupling technologies of different kinds of MGs complement others' demand. Also, the existing work on P2P energy trading only uses single-agent DRL algorithms.

To address the above issues, a new P2P energy trading and energy conversion scheme and a multi-agent (MA) DRL approach are proposed for multiple interconnected MEMGs. The main contributions of this paper are as follows:

1) To the best of our knowledge, this is the first work to consider P2P energy trading, energy conversion and multi-vector energies together in a holistic way. A new P2P energy trading and energy conversion scheme is established for interconnected residential, commercial and industrial MEMGs. A two-stage problem consisting of P2P energy trading and energy conversion process is formulated as a partially observable Markov decision process (POMDP).
2) A MADRL approach MATD3 is proposed to optimise P2P energy trading and energy conversion policies of MEMGs in real-time. The proposed method combines MADRL framework in [38] with twin delayed deep deterministic policy gradient algorithm (TD3) [39] further to improve the performance of the MA actor-critic algorithm. The original MADRL framework has been modified particularly for our P2P energy trading and energy conversion problem and also for stabilizing the learning process. To our best knowledge, this is the first paper using multi-agent DRL models for P2P energy trading. Our proposed MATD3 approach can be used to choose the optimal actions within continuous action space and enables all the MEMGs to learn their policies simultaneously to achieve the best goal individually.

The remainder of the paper is organised as follows. Section II formulates the problem of P2P energy trading and energy conversion for interconnected residential, commercial and industrial MEMGs. Section III proposes the MATD3 method. Section IV presents a case study to evaluate the effectiveness of the proposed model. Section V draws the conclusion.

II. System Model

A. System Overview

Fig. 1 shows the P2P energy trading paradigm among N MEMGs located in residential, commercial and industrial areas within a local community. These MEMGs can not only trade electricity with the main power grid and buy natural gas as fuel from the external networks but also trade electricity and heat among themselves. We assume there are heat networks built within local areas so that the MGs are more willing to trade heat instead of gas sources. Each MG includes renewable generators, a multi-energy system (MES) and electricity and heat load. The energy flow of the residential, commercial and Industrial MEMGs is illustrated in Fig. 2.

1) Residential MEMG: Solar panels are installed at the residential houses, and the electrical storage system can store any excess electricity. Solar power can be used to produce hydrogen with the help of a water electrolyser. Hydrogen can be converted to electricity and heat using a fuel cell [40] or generate heat using a boiler. Natural gas is a standby fuel to cover the necessary heat demand.

Hydrogen fuel has great potential to be used in the residential sector, because of the versatility to generate electricity,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3124465, IEEE Transactions on Smart Grid

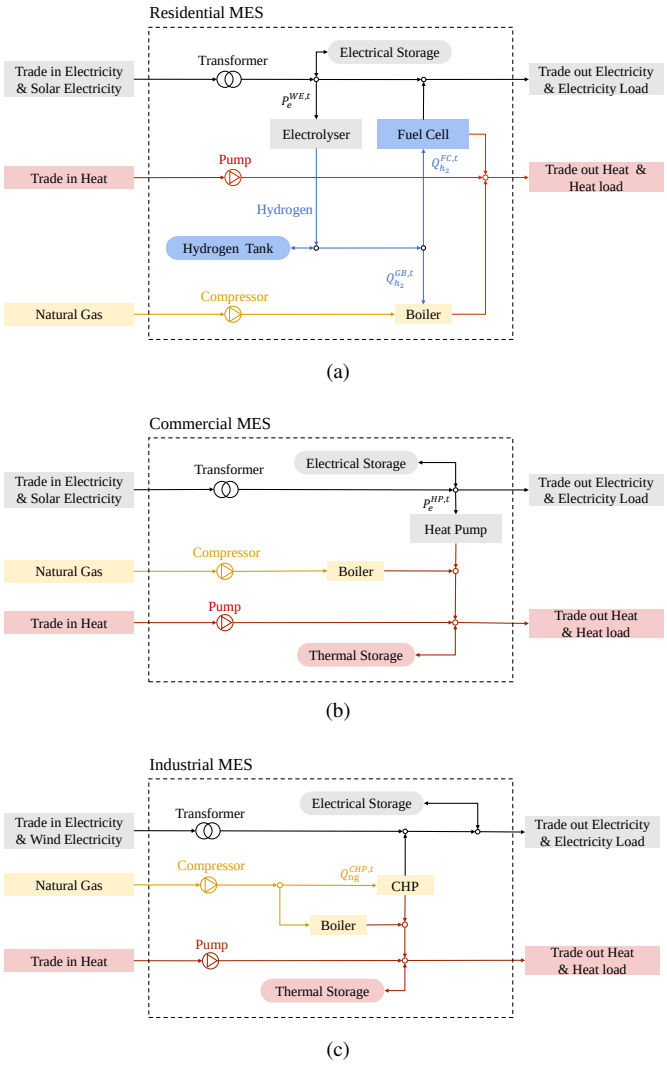


Fig. 2. (a) The energy-flow-diagram of the residential MEMG, (b) The energy-flow-diagram of the commercial MEMG, (c) The energy-flow-diagram of the industrial MEMG.

heat and serve as energy source for vehicles with low carbon emissions in the future, even though the residential usage of hydrogen is still on the initial stage. Corporate investment [41] and hydrogen infrastructure development [42] are likely to reduce hydrogen cost. Governments have invested in clean and safe hydrogen heating systems in residential building [43]. E.g., A project has been approved in Scotland to heat homes with 100 per cent green hydrogen [44]. These investments can further accelerate residential usage of hydrogen.

The water electrolyzers are used as a backup system to provide extra electricity and heat when necessary, because hydrogen generated by a water electrolyzer can be stored in a hydrogen tank indefinitely until needed. Other energy storage systems, such as batteries, lose energy over time, and have to be recharged periodically [45]. The cost of electrolyzers is continually declining [46], and the efficiency of electrolyzers is improving [47].

2) Commercial MEMG: For commercial MG, the primary heat load is for space heating. Therefore, the heat pump, converting electricity into heat, is a better choice for space heating than the gas boiler, which is only a standby resource. The MG has solar panels installed on the their buildings and a natural gas supply as well. In addition, the MG has a thermal storage system to store excess thermal energy for later use.

3) Industrial MEMG: A combined heat and power (CHP) generator provides electrical and thermal energy simultaneously to meet electricity and heat demand, which is used to improve energy efficiency. There are also wind turbines on-site to provide additional electricity generation. Moreover, this MG is equipped with electrical and thermal storage systems.

B. POMDP & System Objective

The P2P energy trading and energy conversion problem is formulated as a POMDP to minimise the operation cost of each MEMG. A POMDP consists of a set of states, a set of observations, a set of actions, a set of reward functions and a set of state transition functions.

1) System States and MG Observations: The system states, $\boldsymbol{s}^t = \{s_1^t, \ldots, s_N^t\}$, describe the configurations of all MGs at time t. The system state of MG i at time t is defined as $s_i^t = [G_i^t, D_{e,i}^t, D_{h,i}^t, E_i^t, \rho_{P2P,e}^t]$, where G_i^t is the renewable generation of MG i between time t and time $t+1$, $D_{e,i}^t$ and $D_{h,i}^t$ represents the electricity and heat demand of MG i between time t and time $t+1$, E_i^t includes electrical storage energy level $E_{e,i}^t$, level of hydrogen stored in the tank $E_{h_2,i}^t$ and thermal storage energy level $E_{th,i}^t$ at time t, and $\rho_{P2P,e}^t$ is the P2P electricity price at time t. The natural gas price is not considered in the system state, since it is fixed within a month. Since the MGs can not obtain the true generation and demand at the beginning of each time slot, they need to forecast their generation and demand. Random Gaussian noise is added into the true states to represent the estimated generation and demand (i.e., the observation values), which can effectively represent the uncertainty of the differences between actual states and estimations [48]. The observation of MG i at time t is defined as $o_i^t = [\hat{G}_i^t, \hat{D}_{e,i}^t, \hat{D}_{h,i}^t, E_i^t, \rho_{P2P,e}^t]$, where the hat symbol indicates that the variable is an estimation of the true system state.

2) MG Actions: The system actions, $\boldsymbol{a}^t = \{a_1^t, \ldots, a_N^t\}$, describe the actions of all MGs at time t. The actions of MG i at time t is defined as $a_i^t = [x_i^t, y_i^t]$, where x_i^t are P2P energy trading actions and y_i^t are energy conversion actions. These actions will be described in detail in Subsection II-C. Each MG will choose actions based on their observations.

3) Reward Functions: The system reward functions, $\boldsymbol{r}^t = \{r_1^t, \ldots, r_N^t\}$, describe the reward functions of all MGs at time t. The reward functions can be used to calculate the MGs' revenue (cost is treated as negative revenue) after taking actions $\boldsymbol{a}^t$ and then evaluate the MGs to choose better policies. The reward function of MG i at time t is formulated as

$$r_i^t = r_{P2P,i}^t - C_{eco,i}^t - C_{pen,i}^t - C_{th,i}^t - C_{carbon,i}^t, \qquad (1)$$

which includes P2P energy trading profit $r_{P2P,i}^t$, economic cost $C_{eco,i}^t$, electricity penalty $C_{pen,i}^t$, discomfort cost $C_{th,i}^t$ and environmental cost $C_{carbon,i}^t$ at time t.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3124465, IEEE Transactions on Smart Grid

The P2P energy trading profit is described as

$$r^t_{P2P,i} = \sum_{j=1,j\neq i}^{N} \sum_{u\in U} z^t_{ij,u} \times \left(I_{(z_{ij,u}\leq 0)} \rho^{-,t}_{P2P,u} - I_{(z_{ij,u}>0)} \rho^{+,t}_{P2P,u} \right), \quad (2)$$

where $U = \{e, h\}$ includes electricity (denoted e) and heat (denoted h). For each u, the amount of P2P energy traded is $z^t_{ij,u}$, and $\rho^{-,t}_{P2P,u}$ and $\rho^{+,t}_{P2P,u}$ represent the selling and buying price at time t, respectively.

The economic cost consists of wholesale electricity cost and natural gas cost, which is expressed as

$$C^t_{eco,i} = z^t_{ii,e} \left(I_{(z_{ii,e}>0)} \rho^{+,t}_{grid,e} - I_{(z_{ii,e}\leq 0)} \rho^{-,t}_{grid,e} \right) + z^t_{ii,h} \rho_{gas}, \quad (3)$$

where $z^t_{ii,e} > 0$ represents buying electricity from the external network (i.e., the main grid) in the wholesale market by MG i at time t, $z^t_{ii,e} \leq 0$ means MG i sells electricity in the wholesale market at time t and $z^t_{ii,h}$ denotes the amount of natural gas bought by MG i from the external network at time t; $\rho^{+,t}_{grid,e}$, $\rho^{-,t}_{grid,e}$ and ρ_{gas} refer to as the buying and selling price offered by the main grid and natural gas price at time t, respectively. In this paper, the relationship between P2P electricity prices and electricity prices of the main grid is limited as

$$\rho^{-,t}_{grid,e} \ll \rho^{-,t}_{P2P,e} \approx \rho^{+,t}_{P2P,e} \ll \rho^{+,t}_{grid,e}. \quad (4)$$

In the electricity wholesale market, the price that the MGs buy from the main grid is usually higher than the price that MGs sell to the main grid, since there are transaction costs due to transmission loss [49]. The higher the transaction costs, the larger the difference between the buying price and the selling price [49]. For the P2P trading market, the selling price is set to be the same as the buying price, since the transaction costs are negligible within a local distribution network [50]. The P2P electricity price is set between the main grid buying and selling prices to encourage P2P energy trading. We assume that the MGs decide an agreed P2P price for all of the MGs, and then negotiate the amount of electricity traded among themselves. This method has been used in [35], [51]–[53]. This assumption is made because the combined dynamics of energy trading, energy conversion and multi-vector energies are considered as a whole. The P2P electricity price is set as

$$\rho^t_{P2P,e} = \alpha_{P2P}(\rho^{+,t}_{grid,e} - \rho^{-,t}_{grid,e}) + \rho^{-,t}_{grid,e}, \quad (5)$$

where $\alpha_{P2P} \in (0, 1)$ is the price coefficient.

The electricity penalty happens when the electricity load supply is short between time t and time $t+1$ [54], which is shown in (6). The discomfort cost occurs when the thermal demand of local consumers are not met [11], which is shown in (7).

$$C^t_{pen,i} = \alpha_e \left(D^t_{e,i} - L^t_{e,i} \right), \quad (6)$$

$$C^t_{th,i} = \alpha_h \left(D^t_{h,i} - L^t_{h,i} \right)^2 + \alpha_h \left(D^t_{h,i} - L^t_{h,i} \right), \quad (7)$$

where $L^t_{e,i}$ is real electricity load supplied by MG i between time t and time $t+1$ and α_e represents penalty coefficient; $L^t_{h,i}$ is the heat load of MG i between time t and time $t+1$ which the consumers are actually provided with and α_h is the heat sensitivity coefficient. In this model, the renewable generation curtailment is considered since electricity consumption and generation needs to be balanced. However, extra network charges of renewable generation curtailment are not considered. Therefore the penalty only occurs when $D^t_{e,i} > L^t_{e,i}$. The penalty terms and the penalty coefficients have been designed based on [54] to obtain good performance of the proposed method [55].

The environmental cost is the economic penalty caused by the CO_2 emissions from the natural gas combustion and electricity bought from the main grid [56], expressed as

$$C^t_{carbon,i} = \alpha^{CO_2} \left(\beta^{gas} z^t_{ii,h} + \beta^e z^t_{ii,e} \right), \quad (8)$$

where β^{gas} and β^e denote carbon intensity (CI) which are the emission rate of CO_2 related to the natural gas combustion and the bought net electricity. The carbon tax price denoted α^{CO_2} converts the carbon emissions into economic penalty.

4) State Transition Functions: After executing the system actions $\boldsymbol{a}^t$, the system states $\boldsymbol{s}^t$ will transfer to $\boldsymbol{s}^{t+1}$ based on the state transition functions. The transition functions of the energy level of storage are shown in Subsection II-D2. However, the transition functions of the renewable generation and energy load are not available. We will use our proposed DRL algorithm to learn from the real-world datasets without knowing the complete state transition functions of the system.

5) System Problem: The system problem for MG i is to find optimal policy $\pi(x^t_i, y^t_i | \hat{G}^t_i, \hat{D}^t_{e,i}, \hat{D}^t_{h,i}, E^t_i, \rho^t_{P2P,e})$ at time t to maximise its expected total rewards (same as minimising the expected total operation cost) which summarises discounted future rewards over the time horizon T, formulated as

$$\mathbf{P1}: \max_{\pi} R^t_{i_\pi} = \mathbb{E}\left[\sum_{\tau=0}^{T} \gamma^\tau r^{t+\tau+1}_i \right], \quad (9)$$

where γ is the discount factor.

C. Two-stage System Process

The MEMGs' operation process contains two stages: P2P energy trading and energy conversion stages. We assume the external P2P energy trading take place in an hour-ahead P2P energy market, in which each MG can buy or sell the desired energy for the next hour. After the real energy trading deals have been made, the process moves to the internal energy conversion stage.

1) P2P Energy Trading Stage: Before the trading begins, MG i uses its observations o^t_i and its policy to choose trading actions to seek for possible deals. The trading actions of MG i at time t is denoted as $x^t_i = \left[x^t_{ij}\right]_{1\leq j\neq i\leq N} = [x^t_{i1}, x^t_{i2}, \dots x^t_{iN}]$, where $x^t_{ij} = [x^t_{ij,e}, x^t_{ij,h}]$ are the intended amounts of energy trading (including electricity and heat) between MG i and MG j at time t. If $x^t_{ij} > 0$, which means MG i wants to buy energy from MG j; if $x^t_{ij} < 0$, which means MG i wants to sell energy to MG j. MGs often have conflicting trading intentions, e.g., $x^t_{ij} \times x^t_{ji} > 0$. Therefore, trading negotiations have been made resulting in real deals of energy trading $z^t_i = \left[z^t_{ij}\right]_{1\leq j\neq i\leq N} = [z^t_{i1}, z^t_{i2}, \dots z^t_{iN}]$, where

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3124465, IEEE Transactions on Smart Grid

$z_{ij}^t > 0$ means MG i buys energy from MG j; $z_{ij}^t < 0$ means MG i sells energy to MG j. MGs only have a deal when one of them wants to sell energy and another wants to buy energy. Note that the actual energy trading might not be the same as the intention, and therefore, MGs need to trade energy with external networks to realise their intended trading actions. The amount of energy traded with external networks at time t are denoted as z_{ii}^t. The actual amount of P2P energy trading of MG i are shown as

$$z_{ij}^t = \begin{cases} \frac{x_{ij}^t}{\left|x_{ij}^t\right|} \min\left(\left|x_{ij}^t\right|, \left|x_{ji}^t\right|\right), & x_{ij}^t x_{ji}^t < 0, \forall i \neq j, \\ 0, & x_{ij}^t x_{ji}^t \geq 0, \forall i \neq j, \\ \sum\limits_{j'=1, j' \neq i}^{N} x_{ij'}^t - \sum\limits_{j'=1, j' \neq i}^{N} z_{ij'}^t, & \forall i = j. \end{cases} \tag{10}$$

2) Energy Conversion Stage: The complexity of an MES is due to the flexibility of exchanging different energy vectors, achieved by managing the energy converters such as fuel cell, heat pump and CHP. The MG needs to consider all available information, including the energy trading results. For residential MG i, the conversion actions y_i^t consist of inflow vector of the water electrolyser $P_e^{WE,t}$, inflow vector of the fuel cell $Q_{h_2}^{FC,t}$ and inflow hydrogen of the boiler $Q_{h_2}^{GB,t}$, as shown in Fig. 2a. For commercial MG i, the conversion action y_i^t is the inflow vector of the heat pump $P_e^{HP,t}$ as shown in Fig. 2b. For industrial MG i, the conversion action y_i^t is the inflow vector of the CHP $Q_{ng}^{CHP,t}$, as shown in Fig. 2c.

D. Physical Constraints

1) Energy Converters Constraints: Energy convert functions are used to show the energy conversion mapping from inflow energy to outflow energy through the energy converters [56]. The convert functions are defined as follows,

$$P_e^{FC,t} = \eta_e^{FC} \times Q_{h_2}^{FC,t}, \tag{11}$$
$$Q_h^{FC,t} = \eta_h^{FC} \times Q_{h_2}^{FC,t}, \tag{12}$$
$$Q_{h_2}^{WE,t} = \eta^{WE} \times P_e^{WE,t}, \tag{13}$$
$$Q_h^{GB,t} = \eta_{h_2}^{GB} \times Q_{h_2}^{GB,t}, \tag{14}$$
$$Q_h^{GB,t} = \eta_{ng}^{GB} \times Q_{ng}^{GB,t}, \tag{15}$$
$$Q_h^{HP,t} = \eta^{HP} \times P_e^{HP,t}, \tag{16}$$
$$P_e^{CHP,t} = \eta_e^{CHP} \times Q_{ng}^{CHP,t}, \tag{17}$$
$$Q_h^{CHP,t} = \eta_h^{CHP} \times Q_{ng}^{CHP,t}. \tag{18}$$

Equations (11)-(13) denote the convert functions of fuel cell and water electrolyser in the residential MGs, where $Q_{h_2}^{FC,t}$, $P_e^{FC,t}$ and $Q_h^{FC,t}$ denote hydrogen inflow, electricity outflow and heat outflow of the fuel cell at time t; η_e^{FC} and η_h^{FC} represent the electricity and heat conversion coefficient of the fuel cell; $P_e^{WE,t}$, $Q_{h_2}^{WE,t}$ and η^{WE} denote electricity inflow, hydrogen outflow and conversion coefficient of the water electrolyser at time t. Equations (14)-(15) refer to the convert functions of the gas boiler with hydrogen or natural gas input, where $Q_{h_2}^{GB,t}$, $Q_{ng}^{GB,t}$ and $Q_h^{GB,t}$ denote hydrogen inflow, natural gas inflow and heat outflow of the gas boiler at time t; $\eta_{h_2}^{GB}$ and η_{ng}^{GB} represent the hydrogen and natural gas conversion coefficient of the gas boiler. The energy conversion process of heat pump is denoted in (16), where $P_e^{HP,t}$, $Q_h^{HP,t}$ and η^{HP} represent electricity inflow, heat outflow and conversion coefficient of the heat pump at time t. The convert functions of CHP are denoted in (17)-(18), where $Q_{ng}^{CHP,t}$, $P_e^{CHP,t}$ and $Q_h^{CHP,t}$ denote natural gas inflow, electricity outflow and heat outflow of the CHP at time t; η_e^{CHP} and η_h^{CHP} represent electricity and heat conversion coefficient of the CHP.

2) Energy Storage Systems Constraints: The dynamic energy level of the storage systems depends on their inherent constraints, shown as follows,

$$E_e^{t+1} = \eta_e^{ES} E_e^t + P_e^{ES,t} \left(I_{(P_e>0)} \eta_{e,ch}^{ES} - \frac{I_{(P_e \leq 0)}}{\eta_{e,dis}^{ES}} \right) \Delta t, \tag{19}$$
$$P_e^{ES^{min}} \leq P_e^{ES,t} \leq P_e^{ES^{max}}, \tag{20}$$
$$0 \leq E_e^{t+1} \leq B_e, \tag{21}$$
$$E_{th}^{t+1} = \eta_{th}^{TS} E_{th}^t + Q_h^{TS,t} \left(I_{(Q_h>0)} \eta_{th,in}^{TS} - \frac{I_{(Q_h \leq 0)}}{\eta_{th,out}^{TS}} \right) \Delta t, \tag{22}$$
$$Q_h^{TS^{min}} \leq Q_h^{TS,t} \leq Q_h^{TS^{max}}, \tag{23}$$
$$0 \leq E_{th}^{t+1} \leq B_{th}, \tag{24}$$
$$E_{h_2}^{t+1} = \eta_{h_2}^{HT} E_{h_2}^t + Q_{h_2}^{HT,t} \left(I_{\left(Q_{h_2}>0\right)} \eta_{h_2,in}^{HT} - \frac{I_{\left(Q_{h_2} \leq 0\right)}}{\eta_{h_2,out}^{HT}} \right) \Delta t, \tag{25}$$
$$Q_{h_2}^{HT^{min}} \leq Q_{h_2}^{HT}(t) \leq Q_{h_2}^{HT^{max}}, \tag{26}$$
$$0 \leq E_{h_2}^{t+1} \leq B_{h_2}. \tag{27}$$

Equations (19)-(21) show the characteristics of the electrical storage system. Equation (19) explains the transition function of energy level of electrical storage, where $P_e^{ES,t}$ is the charging or discharging power of electrical storage; η_e^{ES}, $\eta_{e,ch}^{ES}$ and $\eta_{e,dis}^{ES}$ represents the self decay rate, charging coefficient and discharging coefficient of electrical storage. Equation (20) shows the limits the electrical power when charging or discharging the electrical storage and (21) is the capacity limitation, where B_e is the capacity of electrical storage. Equations (22)-(24) indicate the limits of the thermal storage system. Equation (22) shows the transition function of energy level of thermal storage, where $Q_h^{TS,t}$ is the inflow or outflow heat power of thermal storage; η_{th}^{TS}, $\eta_{th,in}^{TS}$ and $\eta_{th,out}^{TS}$ represent the self decay rate, inflow coefficient and outflow coefficient of thermal storage. Equation (23) limits the inflow and outflow heat of the thermal storage system and the energy level of thermal storage is bounded by (24), where B_{th} is the capacity of thermal storage. Similarly, the transition function of energy level of hydrogen tank is formulated in (25), where $Q_{h_2}^{HT,t}$ is the inflow or outflow hydrogen of hydrogen tank; $\eta_{h_2}^{HT}$, $\eta_{h_2,in}^{HT}$ and $\eta_{h_2,out}^{HT}$ represent the self decay rate, inflow coefficient and outflow coefficient of hydrogen tank. The hydrogen gas flow limitation and hydrogen tank capacity limitation are described in (26)-(27), where B_{h_2} is the capacity of hydrogen tank. If the power of an energy storage system is greater than 0, it means charging the storage or the energy is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3124465, IEEE Transactions on Smart Grid

flowing into the storage. If an energy storage system's power is less than 0, it means discharging the storage.

3) Energy Balance Constraints: For the energy networks of an MEMG working correctly, the MG must balance the energy generation and consumption between time t and time $t+1$. The energy balance constraints for residential, commercial and industrial MEMGs are formulated as follows,

$$z_{e,i}^{t}+G_{i}^{t}+P_{e,i}^{FC,t}\Delta t=P_{e,i}^{ES,t}\Delta t+P_{e,i}^{WE,t}\Delta t+D_{e,i}^{t}, \tag{28}$$

$$z_{h,i}^{t}+Q_{h,i}^{GB,t}\Delta t=D_{h,i}^{t}, \tag{29}$$

$$Q_{h_2,i}^{WE,t}\Delta t=Q_{h_2,i}^{HT,t}\Delta t+Q_{h_2,i}^{FC,t}\Delta t+Q_{h_2,i}^{GB,t}\Delta t, \tag{30}$$

$$z_{e,i}^{t}+G_{i}^{t}=P_{e,i}^{ES,t}\Delta t+P_{e,i}^{HP,t}\Delta t+D_{e,i}^{t}, \tag{31}$$

$$z_{h,i}^{t}+Q_{h,i}^{HP,t}\Delta t+Q_{h,i}^{GB,t}\Delta t=Q_{h,i}^{TS,t}\Delta t+D_{h,i}^{t}, \tag{32}$$

$$z_{e,i}^{t}+G_{i}^{t}+P_{e,i}^{CHP,t}\Delta t=P_{e,i}^{ES,t}\Delta t+D_{e,i}^{t}, \tag{33}$$

$$z_{h,i}^{t}+Q_{h,i}^{CHP,t}\Delta t+Q_{h,i}^{GB,t}\Delta t=Q_{h,i}^{TS,t}\Delta t+D_{h,i}^{t}. \tag{34}$$

Equations (28)-(30) indicate that residential MG i must balance the electricity, heat and hydrogen energy, respectively. The electricity and heat networks of commercial MG i are constrained in (31)-(32). Equations (33)-(34) describe the energy balance equations of industrial MG i with electricity and heat distribution networks.

III. Proposed Multi-Agent Deep Reinforcement Learning Based Approach

A MATD3 approach is proposed to solve the P2P energy trading and energy conversion problem formulated in (9). TD3 is a model-free, off-policy actor-critic algorithm which uses DNNs to learn policies in high-dimensional, continuous state-action spaces. The MATD3 approach adopts the form of centralised critics to ease training and decentralised actors to ensure all MEMGs are operating independently.

A. Twin Delayed Deep Deterministic Policy Gradient Algorithm

TD3 was proposed to solve the overestimation and high variance problems lied in deep Q-learning [57] and deep deterministic policy gradient (DDPG) [36] algorithms [39]. To solve the overestimation problem, TD3 adopts the idea of double Q-learning [58]. In TD3, the critic consists of two Q-networks (Q_{θ_1} and Q_{θ_2}) and their target networks ($Q_{\theta_1'}$ and $Q_{\theta_2'}$), and the actor is formed by a deterministic policy network π_ϕ and its target network $\pi_{\phi'}$. The target networks are time-delayed copies of their Q-networks, which greatly improve stability in learning [39]. To update the TD3 networks, the Q networks in critic minimise the loss via (35), where p_π is the state distribution, π and R are distribution of the policy and reward function, and y^t is the target value. The deterministic policy network in actor is updated using sampled policy gradient which is shown in (36), i.e.,

$$\mathcal{L}\left(\theta\right)=\mathbb{E}_{s^t\sim p_\pi,a^t\sim\pi,r^t\sim R}\left[\left(Q_\theta\left(s^t,a^t\right)-y^t\right)^2\right], \tag{35}$$

$$\nabla_\phi J\approx\mathbb{E}_{s^t\sim p_\pi}\left[\nabla_a Q_\theta\left(s,a\right)|_{s=s^t,a=\pi_\phi(s^t)}\nabla_\phi\pi_\phi\left(s\right)|_{s=s^t}\right], \tag{36}$$

where

$$y^t=r^t+\gamma\min_{j\in\{1,2\}}Q_{\theta_j'}\left(s^{t+1},\tilde{a}^{t+1}\right), \tag{37}$$

$$\tilde{a}^{t+1}=\pi_{\phi'}(s^{t+1})+\tilde{\epsilon},\ \tilde{\epsilon}\sim\mathrm{clip}(\mathcal{N}(0,\tilde{\sigma}),-c,c). \tag{38}$$

The critic will choose the minimum target value between the two target Q-networks as in (37), where $\tilde{a}^{t+1}$ is the clipped target action. The minimum operation results in low-variance value estimations and makes the algorithm more stable. To address the high variance problem, TD3 updates the policy networks once every several Q-value updates. By sufficiently delaying the policy updates, TD3 allows the Q-network to produce lower Q-values, and hence less chance of a mistake being exploited. TD3 algorithm also adds target policy noise as shown in (38) when forming the target, where $\tilde{\epsilon}$ is the clipped Gaussian noise and c is the edge value. This target policy regularisation technique will smooth but keep close to the original target action, which helps the algorithm remain stable and converge fast in the stochastic domain.

B. Multi-Agent Twin Delayed Deep Deterministic Policy Gradient Approach

As our P2P energy trading and energy conversion model is in an MA environment (each MG is an agent), a naive approach is to directly apply TD3 algorithm to learn each agent's policy independently. However, the environment is no longer static from the view of each agent since the agents are learning their own policy independently. $P\left(\boldsymbol{s}^{t+1}|\boldsymbol{s}^t,\boldsymbol{a}^t,\pi_1,\ldots,\pi_N\right)\neq P\left(\boldsymbol{s}^{t+1}|\boldsymbol{s}^t,\boldsymbol{a}^t,\pi_1',\ldots,\pi_N'\right)$ for any $\boldsymbol{\pi}\neq\boldsymbol{\pi}'$, which violates the Markov assumption. Therefore, this naive approach has difficulty in learning good policies. Instead, we adopted the concept of centralised training with decentralised execution in [38], where the training of critic take consideration of the actions and observations of all the agents in the environment but the actor of each agent choose actions only based on its own observations. The centralised Q-value function of MG i, $Q_{\theta_i}\left(o_1^t,\ldots,o_N^t,a_1^t,\ldots,a_N^t\right)$, takes observations and actions of all MGs as inputs instead of only its own. The critics are learned by their rewards, where the reward functions can be different from each other, allowing both competitive and collaborative multi-agent settings. The main reason for using a centralised critic is that the environment is stationary if all the actions of the agents are known, where $P\left(\boldsymbol{s}^{t+1}|\boldsymbol{s}^t,a_1^t,\ldots,a_N^t,\pi_1,\ldots,\pi_N\right)=P\left(\boldsymbol{s}^{t+1}|\boldsymbol{s}^t,a_1^t,\ldots,a_N^t,\pi_1',\ldots,\pi_N'\right)$ even for $\boldsymbol{\pi}\neq\boldsymbol{\pi}'$. The actor works in a decentralised way, to ensure that only local information is used when executing policies.

The centralised Q-value function is updated as

$$\mathcal{L}\left(\theta_i\right)=\mathbb{E}_{\boldsymbol{o}^t,\boldsymbol{a}^t,\boldsymbol{r}^t,\boldsymbol{o}^{t+1}}\left[\left(Q_{\theta_i}\left(o_1^t,\ldots,o_N^t,a_1^t,\ldots,a_N^t\right)-y_i^t\right)^2\right]$$

$$y_i^t=r_i^t+\gamma\min_{j\in\{1,2\}}Q_{\theta_{ij}'}\left(o_1^{t+1},\ldots,o_N^{t+1},\tilde{a}_1^{t+1},\ldots,\tilde{a}_N^{t+1}\right)$$

$$\tilde{a}_i^{t+1}=\pi_{\phi_i'}(o_i^{t+1})+\tilde{\epsilon}_i,\ \tilde{\epsilon}_i\sim\mathrm{clip}(\mathcal{N}(0,\tilde{\sigma}_i^2),-c_i,c_i). \tag{39}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3124465, IEEE Transactions on Smart Grid

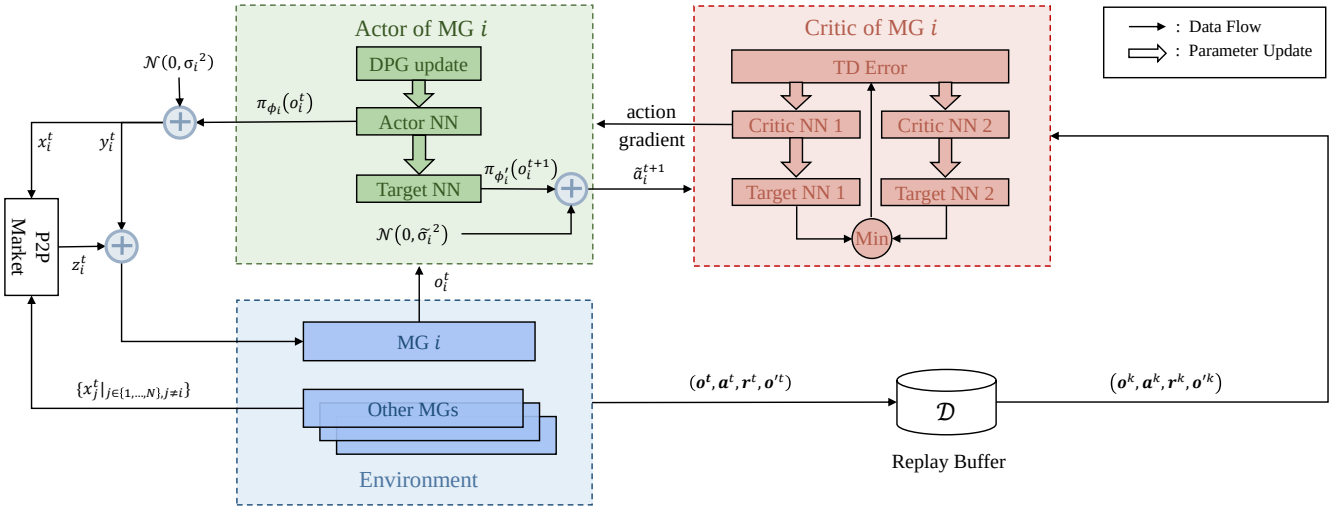


Fig. 3. Information flowchart of the MADRL agent training and execution process. NN stands for neural network.

The gradient of the policy network can then be written as

$$\nabla_{\phi_i} J \approx \mathbb{E}_{\boldsymbol{o}^t,\boldsymbol{a}^t}[\nabla_{a_i} Q_{\theta_i}\left(o_1^t,\ldots,o_N^t,a_1^t,\ldots,a_N^t\right) |_{o_i=o_i^t,a_i=\pi_{\phi_i}(o_i^t)} \nabla_{\phi_i}\pi_{\phi_i}\left(o_i\right)|_{o_i=o_i^t}]. \quad (40)$$

In the implementation, mini-batches are used to train the networks rather than a single transition of data.

C. Implementation of The Proposed Method

The information flow of our proposed approach is illustrated in Fig. 3. For each MG i, it firstly receives its observations o_i^t at time t. The actor of MG i will then choose P2P energy trading actions x_i^t and energy conversion actions y_i^t based on o_i^t and its policy π_{ϕ_i}. A random noise sampled from a Gaussian distribution is added to the actor to increase exploration. During the P2P energy trading stage, MG i will negotiate with other MGs and get real energy trading deals z_i^t. After that, y_i^t and z_i^t are used to operate MG i in the energy conversion stage. MG i will then receive the reward r_i^t and observations of next states o_i^{t+1}. Finally, the transition of observations, actions, rewards and next observations of all MGs $(\boldsymbol{o}^t,\boldsymbol{a}^t,\boldsymbol{r}^t,\boldsymbol{o}'^t)$ will be stored in the replay buffer $\mathcal{D}$, where $\boldsymbol{o}^t=\{o_1^t,\ldots,o_N^t\},\boldsymbol{a}^t=\{a_1^t,\ldots,a_N^t\},\boldsymbol{r}^t=\{r_1^t,\ldots,r_N^t\},\boldsymbol{o}'^t=\{o_1^{t+1},\ldots,o_N^{t+1}\}$. For the centralised training, each MG will sample a random mini-batch of size m $(\boldsymbol{o}^k,\boldsymbol{a}^k,\boldsymbol{r}^k,\boldsymbol{o}'^k)$ from $\mathcal{D}$. The parameters of the critic θ_i will be updated by minimising the sample loss via (39), and the actor will be updated using sampled policy gradient according to (40). The target networks of MG i will then be updated using the following equations

$$\theta_{i1}' \leftarrow \tau\theta_{i1}+(1-\tau)\theta_{i1}', \quad (41)$$

$$\theta_{i2}' \leftarrow \tau\theta_{i2}+(1-\tau)\theta_{i2}', \quad (42)$$

$$\phi_i' \leftarrow \tau\phi_i+(1-\tau)\phi_i', \quad (43)$$

where $\tau \ll 1$ is the target update parameter. Thus, the target values change slowly which greatly improves the stability of learning. Each episode contains T time steps, and the training process repeats M times to ensure the algorithm converges. The proposed MATD3 algorithm is shown in Algorithm 1.

Algorithm 1: MATD3-based P2P Energy Trading and Energy Conversion in Interconnected MEMGs

Initialize $\gamma,\tau,\theta_{i1},\theta_{i2},\phi_i$ and replay buffer $\mathcal{D}$
for *episode = 1 to M* **do**
  Initialize random process $\mathcal{N}$ for action exploration
  for *t = 1 to T* **do**
    For each MG i, forecast $\hat{G}_i^t,\widehat{D}_{e,i}^t,\widehat{D}_{h,i}^t$, and observe E_i^t and $\rho_{P2P,e}^t$ to form o_i^t
    Choose P2P energy trading actions x_i^t and energy conversion actions y_i^t w.r.t. the current policy π_{ϕ_i}
    P2P energy trade with other MGs, and get the real energy trading deals z_i^t via (10)
    Convert energy based on z_i^t and y_i^t, and get reward r_i^t and new observations o_i^{t+1}
    Store $(\boldsymbol{o}^t,\boldsymbol{a}^t,\boldsymbol{r}^t,\boldsymbol{o}'^t)$ of all MGs in $\mathcal{D}$
    $\boldsymbol{o}^t \leftarrow \boldsymbol{o}'^t$
    for *MG i = 1 to N* **do**
      sample a random mini-batch of size m $(\boldsymbol{o}^k,\boldsymbol{a}^k,\boldsymbol{r}^k,\boldsymbol{o}'^k)$ from $\mathcal{D}$
      Update critic parameters θ_{i1} and θ_{i2} by minimising the loss via (39)
      Update actor parameter ϕ_i every two critic updates via (40)
    end
    Update target network parameters for each MG i via (41)-(43)
  end
end

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3124465, IEEE Transactions on Smart Grid

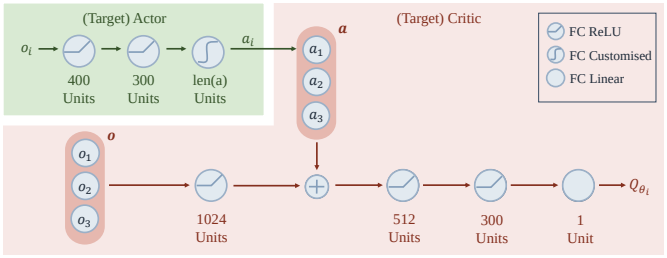


Fig. 4. The neural network architecture of (target) actor and (target) critic for each agent.

TABLE I
EFFICIENCIES AND CAPACITIES OF ENERGY CONVERTERS

DER	Efficiency	Capacity (kW/kWh)	Location
WE	$\eta^{WE} = 80\%$	$\bar{P}_e^{WE,t} = 150$	MG 1
FC	$\eta_e^{FC} = 30\%$ $\eta_h^{FC} = 55\%$ [67]	$\bar{Q}_{h_2}^{FC,t} = 330$	MG 1
GB	$\eta^{GB} = 90\%$	$\bar{Q}_{ng}^{GB,t} = 1500$	MG 1,2 & 3
HP	$\eta^{HP} = 300\%$	$\bar{P}_e^{HP,t} = 150$	MG 2
CHP	$\eta_e^{CHP} = 45\%$ $\eta_h^{CHP} = 40\%$	$\bar{Q}_{ng}^{CHP,t} = 900$	MG 3

D. Modifications to The Original MADRL Framework

The original MADRL framework has been modified particularly for the P2P energy trading and energy conversion problem and for stabilising the learning process. The modifications of the original MADRL framework include as follows:

1) TD3 Agent Customisation: In the original MADRL framework, the activation function of the output layer in the actor networks is a hyperbolic tangent or sigmoid function. In the proposed MATD3 method, for each MG, the activation function of the output layer in the actor networks is customized to provide the requisite output shape of the actor in terms of energy trading and energy conversion actions, since the range of values for the energy trading actions and energy conversion actions can be very different.

2) State/Observation Normalization: For each MG, the components of the observation vector have different magnitudes. Normalizing the observations can prevent bias and speed up the training process [59].

3) Reward Scaling: Reward values obtained from the reward function cannot be used directly by the agent, since the learning process might not be stable due to the wide range of reward values [60]. Therefore, the reward is sampled from the reward functions to calculate the distribution of the reward, and then the z-score of the new reward (i.e., the standardized reward) can be calculated based on the distribution. This scaling of the reward and setting of a lower bound of the z-score make our learning process stable.

4) Network Architecture: The neural network architecture of (target) actor and (target) critic for each agent are presented in Fig. 4. The fully connected (FC) layers use the Rectified Linear Unit (ReLU) or the customised function as activation functions. Compared to the original MADRL framework, a hidden layer has been added to the observations before concatenating with the actions.

IV. NUMERICAL SIMULATION

A. Case Study Setup

The proposed MATD3 approach is simulated in a 3-MEMG model including a residential MEMG, a commercial MEMG and an industrial MEMG. Three real-world datasets containing renewable generation and energy demand data at 1-hour resolution are used to train our model, where MG 1 uses data [61] from residential households located in Mueller, Austin, Texas; MG 2 uses data [62] from a commercial data warehouse located in Mueller, Austin, Texas; and MG 3 uses data [63] from a power plant at trial site Aachen/Cologne, Germany[1]. The parameters of energy converters are given in Table I. The electricity price offered by the main grid follows the hourly locational marginal pricing from ISO New England Inc. [64] and the natural gas prices follow the monthly Natural Gas Industrial Price from US Energy Information Administration [65]. Also, the carbon tax price α^{CO_2} is set to 0.0316 \$/kg, while the carbon intensities of natural gas and grid electricity are $\beta^{gas} = 0.245$ kg/kWh and $\beta^{e} = 0.683$ kg/kWh, respectively [66].

B. Performance Evaluation

To demonstrate the effectiveness of our proposed scheme and MATD3 algorithm, the following methods are compared:

1) The Rule-based Method: The MGs do not use any energy converters, and they only trade energy with external networks. The rule-based operating policy calculates the difference between the estimated energy demand and generation for the trading time slot, and then sells the surplus electricity or buy the needed energy.

2) SATD3-SEP: The SATD3-SEP method has the same configuration as rule-based one, except that it uses three independent TD3 agents to find the trading actions of each MG with external networks. Therefore, the agents only use their own observations, actions, rewards and next observations to train their critic networks.

3) SATD3: The SATD3 method will use our system model for P2P energy trading and energy conversion. However, the agents for the three MEMGs are independent TD3 agents.

4) MATD3: This is our proposed energy trading and conversion scheme, and our proposed MATD3 method.

The average hourly operation costs of each MG in a typical winter day using each method are illustrated in Fig. 5. The industrial MG has the highest operation cost due to its highest demand. The figure shows that the proposed MATD3 approach can reduce the operation cost and outperform other methods. The MATD3 approach achieves average hourly costs of 4.119, 6.566 and 9.230 US dollars in the residential MG, commercial MG and industrial MG, respectively. In relative terms, MATD3

[1]We cannot find any industrial MG dataset in the same location as the previous two, however, in our setting we assume these three MG are in the same local area.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3124465, IEEE Transactions on Smart Grid

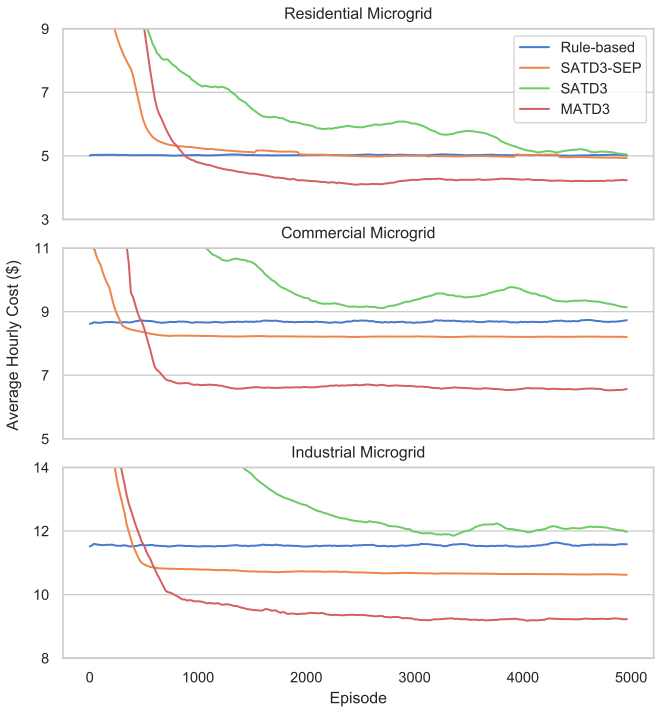


Fig. 5. Average hourly costs of three MGs in a local community for the examined methods. Curves are smoothed uniformly for visual clarity.

TABLE II
COMPUTATIONAL PERFORMANCE OF THE DRL METHODS

Method	SATD3-SEP	SATD3	MATD3
CPU time per episode (s)	1.76	1.54	1.94
Number of episodes	1500	5000[a]	3000
Total CPU time (h)	0.74	2.13[a]	1.62
CPU time at execution (ms)	1.81	1.63	1.75

[a] Failure to converge within 5000 episodes.

reduces the costs 18.2%, 16.5% and 18.1% compared to those of SATD3, SATD3-SEP and rule-based methods for the residential MG; 27.8%, 20% and 24.8% compared to those of SATD3, SATD3-SEP and rule-based methods for the commercial MG; and 23.1%, 13.1% and 20.3% compared to those of SATD3, SATD3-SEP and rule-based methods for the industrial MG. SATD3 did not perform well and failed to converge within 5000 episodes, because directly applying the algorithm into an environment with three interacting MGs violates the Markov assumption. The SATD3-SEP method reduced commercial and industrial MGs' costs compared to the rule-based method. However, SATD3-SEP performs comparably to the rule-based method for the residential MG. These results are due to the fact that the residential MG only has electrical storage, while the commercial and industrial MG have electrical storage and thermal storage.

The computational performance of the compared DRL methods is illustrated in Table II in terms of training and execution. The average CPU time per episode is the highest in MATD3 since the method involves interactions among all three MGs, and each agent trains its critic using the information from all of the MGs. The total CPU time required to reach convergence is shortest in SATD3-SEP because of the independent agents, longer in MATD3 because of the multi-agent setting, and longest in SATD3 (since it fails to reach convergence). For execution, the CPU time of each DRL method is similar and in the order of milliseconds since the policies are directly inferred from the observations by the trained actor networks.

C. Impact of Energy Conversion and P2P Energy Trading

Fig. 6 shows the proportion of each MGs' electricity and heat demand that is met in each hour time slot by renewable generation, energy storage, energy trading (including P2P energy trading and trading with the external network), and energy conversion using our proposed MATD3 approach. This figure also shows how our proposed method was able to reduce the average hourly operation cost of each MEMG by revealing the energy trading and energy conversion decisions made at each time slot. The same data was used as in Subsection IV-B.

The first two columns in Fig. 6 show how the electricity demand and heat demand for each MEMG was met using renewable generation, energy storage, energy trading and energy conversion. Fig. 6a, 6e and 6i reveal that the MGs tend to buy more electricity (labelled in red as ET) when the electricity price is low or the renewable generation is insufficient. The residential MG uses WE to transform purchased surplus electricity to hydrogen stored in the hydrogen tank, e.g., Fig. 6a in hour 22. Later, when needed, FC is used to transform hydrogen to electricity and heat. As shown in Fig. 6f, HP provides a significant amount of heat for the commercial MG and other MGs. For the industrial MG shown in Fig. 6i and 6j, CHP is used to provide electricity and heat when wind electricity is insufficient, or the electricity price is high.

The last two columns in Fig. 6 show the amount of energy traded among three MGs and external networks. P2P energy trading accounts for a considerable proportion of the heat traded. However, the majority of electricity trading is with the main power grid, as all three MGs have insufficient renewable generation to meet their own demand and energy conversion. There is no electricity sold back to the grid in this case, which shows our proposed approach makes appropriate decisions. The last column in Fig. 6 shows the amount of heat that each MEMG trade with others. The commercial MG provides significant heat energy to other MGs via P2P energy trading, which explains why the commercial MG converts much power into heat using HP even when its heat demand is relatively low. These results also show that MGs fulfil their heat demand using P2P heat trading, and only the residential MG needs to buy extra natural gas (Fig. 6d labelled in orange as GB) from the external network.

These results demonstrate that the proposed MATD3 method can utilise energy conversion to flexibly convert and store the energy when needed. It also allows the community to consume heat energy locally with P2P energy trading, and reduce the surplus electricity sent back to the main grid.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3124465, IEEE Transactions on Smart Grid

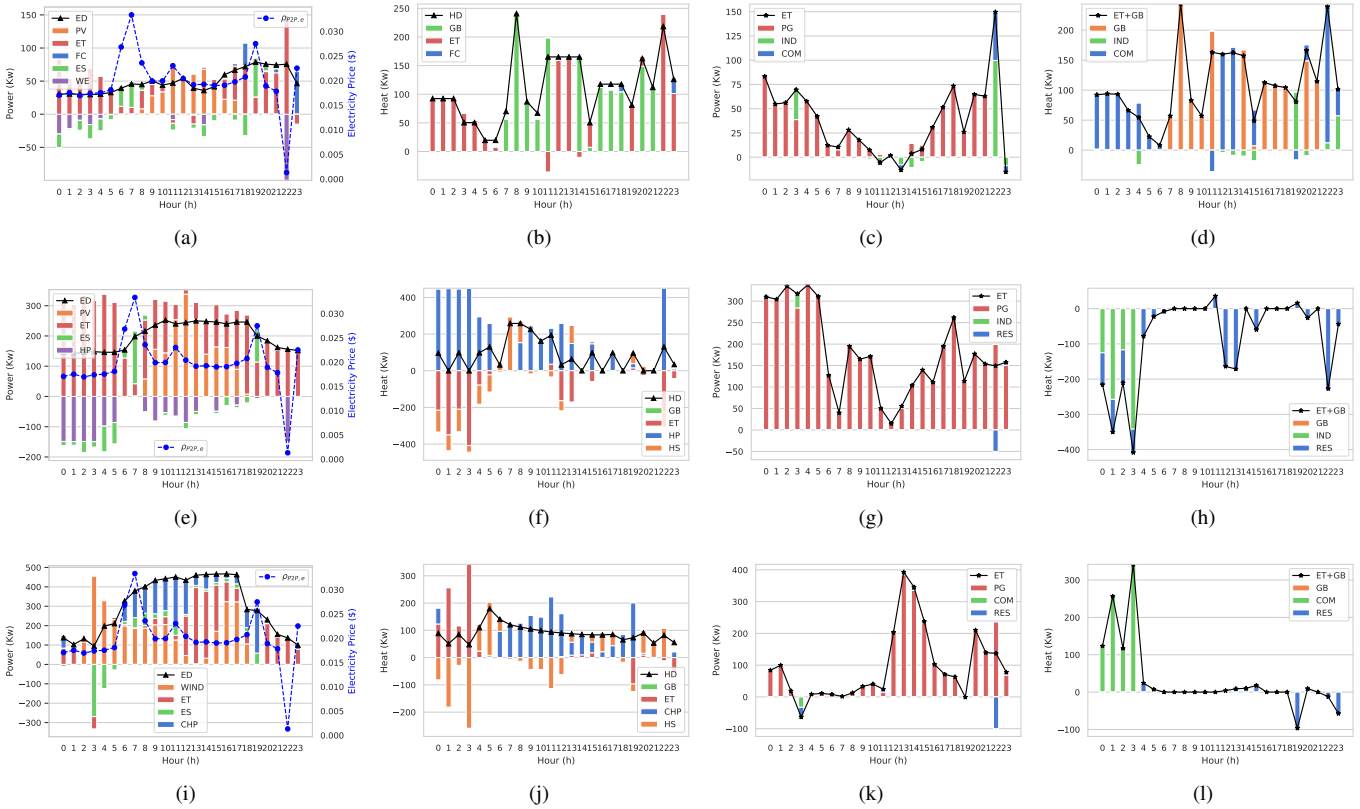


Fig. 6. (a)-(d) show how the electricity demand is met, how the heat demand is met, the amount of electricity traded with each source, and the amount of heat traded with each source respectively for the residential MEMG. (e)-(h) show the corresponding results for the commercial MEMG, and (i)-(l) show them for the industrial MEMG. ED and HD stand for electricity demand and heat demand. PV, ES and HS stand for solar generation, electrical storage and thermal storage. ET and PG mean energy trading and electricity trading with the main grid.

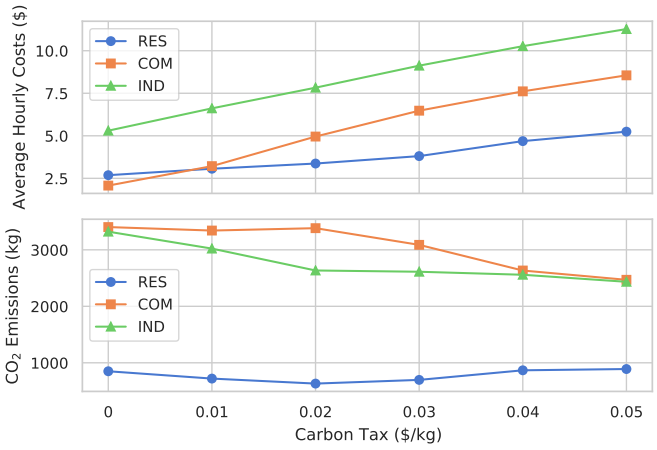


Fig. 7. Average hourly costs and CO_2 emissions at varying carbon tax prices.

D. Effect of Carbon Tax Price on Costs and CO_2 Emissions

The impact of the carbon tax price on the the average hourly costs and CO_2 emissions for each MEMG using our proposed MATD3 method is shown in Fig. 7. When the carbon tax price increases, the average hourly costs of each MG increase due to the increase of the environmental cost. The carbon tax price impact is less on the residential MG, because the amount of energy trading with the external network is smaller compared to the commercial and industrial MGs.

The CO_2 emissions from the commercial MG are stable when the carbon tax price is lower than 0.02 \$/kg, decrease significantly when the carbon tax increases from 0.02 to 0.04 \$/kg, and stay unchanged when the price increases beyond 0.04 \$/kg. The reason is that the commercial MG converts significant amount of electricity purchased from the main grid to heat and trades the heat with other MGs, since the P2P energy trading profit is higher than the environmental cost when carbon tax price is lower than 0.02 \$/kg. When the carbon tax price is above 0.02 \$/kg, the amount of the heat traded with other MGs is reduced, and the CO_2 emissions are reduced until no more heat trading takes place. The figure shows that there is a sharp drop in CO_2 emissions from the industrial MG when the carbon tax price increases from 0 to 0.02 \$/kg, and the CO_2 emissions remain unchanged when the price increases beyond 0.02 \$/kg. As the carbon tax price increases, the industrial MG uses the CHP to meet a higher proportion of its own electricity demand rather than buying electricity from the main grid, and all of the electricity demand was met by the CHP when the carbon tax price is 0.02 \$/kg. Therefore, the CO_2 emissions are reduced at first, and remain unchanged above the carbon tax price of 0.02 \$/kg. The residential MG produces less CO_2 emissions as the carbon tax price increases from 0 to 0.02 \$/kg, and then more CO_2 until the carbon tax price reaches 0.04 \$/kg, and the same

amount of CO_2 above that. This is because the residential MG uses more environmentally friendly approach as the carbon tax price increases at first. However, when the carbon tax price is higher than 0.02 \$/kg, it has to use an increasing amount of natural gas to meet its heat demand because the commercial MG begins to reduce selling heat to other MGs. Once the carbon tax price is higher than 0.04 \$/kg, the demand is met fully by the natural gas in residential MG.

E. Scalability of the proposed approach

The proposed approach can be scaled up for a longer period than a day. If the proposed approach is scaled up to a month or shorter, the time horizon in the system problem shown in (9) needs to be changed from a day to the new period. This method only needs to train the agents once and the results can be reused. If the extended period is longer, e.g., up to one year, the agents have to be trained periodically using the newest collected data and old data. The computational costs of this method are higher than the first method.

V. Conclusion

An external P2P energy trading and internal energy conversion problem was investigated for the interconnected residential, commercial and industrial MEMGs in a local community. The problem was formulated as a POMDP, and a multi-agent deep reinforcement learning approach was proposed to address it. The proposed approach aligns with the nature of P2P energy trading, and can also handle a high-dimensional continuous action space and alleviate overestimation and high variance problems. The case study on three real-world datasets showed that the proposed method significantly reduced all MGs' operation costs. The simulation results also demonstrated that the MATD3 method can utilise energy conversion to flexibly convert and store the energy and allows MGs to consume heat energy locally with P2P energy trading. The simulation results also showed the impact of carbon tax price on the operation cost and CO_2 emissions.

To the best of our knowledge, this work is the first to consider the combined dynamics of energy trading, energy conversion and multi-vector energies (electricity, heat and natural gas) as a whole. More options can be used to match supply with demand, making the system more flexible overall. Increased flexibility provides alternatives to adding additional costly infrastructure to meet demand and supports the inclusion of a higher share of variable renewable energy sources.

References

[1] P. Mancarella, "MES (multi-energy systems): An overview of concepts and evaluation models," *Energy*, vol. 65, pp. 1–17, Feb. 2014.

[2] W. Tushar, C. Yuen, H. Mohsenian-Rad, T. Saha, H. V. Poor, and K. L. Wood, "Transforming energy networks via peer-to-peer energy trading: The potential of game-theoretic approaches," *IEEE Signal Process. Mag.*, vol. 35, no. 4, pp. 90–111, Jul. 2018.

[3] Y. Wang, W. Saad, Z. Han, H. V. Poor, and T. Başar, "A game-theoretic approach to energy trading in the smart grid," *IEEE Trans. Smart Grid*, vol. 5, no. 3, pp. 1439–1450, May 2014.

[4] C. Zhang, J. Wu, Y. Zhou, M. Cheng, and C. Long, "Peer-to-Peer energy trading in a Microgrid," *Appl. Energy*, vol. 220, pp. 1–12, Jun. 2018.

[5] G. El Rahi, S. R. Etesami, W. Saad, N. B. Mandayam, and H. V. Poor, "Managing Price Uncertainty in Prosumer-Centric Energy Trading: A Prospect-Theoretic Stackelberg Game Approach," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 702–713, Jan. 2019.

[6] W. Tushar, T. K. Saha, C. Yuen, T. Morstyn, N. A. Masood, H. Vincent Poor, and R. Bean, "Grid influenced peer-to-peer energy trading," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1407–1418, Mar. 2020.

[7] K. Zhang, S. Troitzsch, S. Hanif, and T. Hamacher, "Coordinated Market Design for Peer-to-Peer Energy Trade and Ancillary Services in Distribution Grids," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 2929–2941, Jul. 2020.

[8] M. Yan, M. Shahidehpour, A. Paaso, L. Zhang, A. Alabdulwahab, and A. Abusorrah, "Distribution Network-Constrained Optimization of Peer-to-Peer Transactive Energy Trading among Multi-Microgrids," *IEEE Trans. Smart Grid*, vol. 12, no. 2, pp. 1033–1047, Mar. 2021.

[9] W. Zhong, S. Xie, K. Xie, Q. Yang, and L. Xie, "Cooperative P2P Energy Trading in Active Distribution Networks: An MILP-Based Nash Bargaining Solution," *IEEE Trans. Smart Grid*, vol. 12, no. 2, pp. 1264–1276, Mar. 2021.

[10] K. Anoh, S. Maharjan, A. Ikpehai, Y. Zhang, and B. Adebisi, "Energy Peer-to-Peer Trading in Virtual Microgrids in Smart Grids: A Game-Theoretic Approach," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1264–1275, Mar. 2020.

[11] S. Cui, Y. W. Wang, Y. Shi, and J. W. Xiao, "A New and Fair Peer-to-Peer Energy Sharing Framework for Energy Buildings," *IEEE Trans. Smart Grid*, vol. 11, no. 5, pp. 3817–3826, Sep. 2020.

[12] Y. Chen, W. Wei, F. Liu, and S. Mei, "A multi-lateral trading model for coupled gas-heat-power energy networks," *Appl. Energy*, vol. 200, pp. 180–191, Aug. 2017.

[13] Y. Chen, W. Wei, F. Liu, E. E. Sauma, and S. Mei, "Energy Trading and Market Equilibrium in Integrated Heat-Power Distribution Systems," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 4080–4094, Jul. 2019.

[14] C. Wang, W. Wei, J. Wang, L. Wu, and Y. Liang, "Equilibrium of Interdependent Gas and Electricity Markets with Marginal Price Based Bilateral Energy Trading," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 4854–4867, Sep. 2018.

[15] P. Jiang, S. Lu, W. Gu, S. Yao, R. Bo, C. Wu, and Z. Wu, "A Two-Stage Game Model for Combined Heat and Power Trading Market," *IEEE Trans. Power Syst.*, vol. 34, no. 1, pp. 506–517, Jan. 2019.

[16] D. Xu, B. Zhou, N. Liu, Q. Wu, N. Voropai, C. Li, and E. Barakhtenko, "Peer-to-Peer Multienergy and Communication Resource Trading for Interconnected Microgrids," *IEEE Trans. Ind. Informatics*, vol. 17, no. 4, pp. 2522–2533, Apr. 2021.

[17] R. Li, W. Wei, S. Mei, Q. Hu, and Q. Wu, "Participation of an Energy Hub in Electricity and Heat Distribution Markets: An MPEC Approach," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3641–3653, Jul. 2019.

[18] J. Guerrero, A. C. Chapman, and G. Verbič, "Decentralized P2P energy trading under network constraints in a low-voltage network," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5163–5173, Sep. 2019.

[19] T. Morstyn, A. Teytelboym, and M. D. McCulloch, "Bilateral contract networks for peer-to-peer energy trading," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 2026–2035, Mar. 2019.

[20] T. Morstyn, A. Teytelboym, C. Hepburn, and M. D. McCulloch, "Integrating P2P Energy Trading with Probabilistic Distribution Locational Marginal Pricing," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3095–3106, Jul. 2020.

[21] Z. Li and T. Ma, "Peer-to-peer electricity trading in grid-connected residential communities with household distributed photovoltaic," *Appl. Energy*, vol. 278, pp. 1–13, Nov. 2020.

[22] F. Zeng, Z. Bie, S. Liu, C. Yan, and G. Li, "Trading Model Combining Electricity, Heating, and Cooling Under Multi-energy Demand Response," *J. Mod. Power Syst. Clean Energy*, vol. 8, no. 1, pp. 133–141, Jan. 2020.

[23] A. Lüth, J. M. Zepter, P. Crespo del Granado, and R. Egging, "Local electricity market designs for peer-to-peer trading: The role of battery flexibility," *Appl. Energy*, vol. 229, pp. 1233–1243, Nov. 2018.

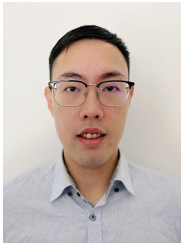
[24] T. Morstyn and M. D. McCulloch, "Multiclass energy management for peer-to-peer energy trading driven by prosumer preferences," *IEEE Trans. Power Syst.*, vol. 34, no. 5, pp. 4005–4014, Sep. 2019.

[25] T. Baroche, P. Pinson, R. L. G. Latimier, and H. B. Ahmed, "Exogenous cost allocation in peer-to-peer electricity markets," *IEEE Trans. Power Syst.*, vol. 34, no. 4, pp. 2553–2564, Jul. 2019.

[26] M. Khorasany, Y. Mishra, and G. Ledwich, "A decentralized bilateral energy trading system for peer-to-peer electricity markets," *IEEE Trans. Ind. Electron.*, vol. 67, no. 6, pp. 4646–4657, Jun. 2020.

[27] Z. Guo, P. Pinson, S. Chen, Q. Yang, and Z. Yang, "Chance-constrained peer-to-peer joint energy and reserve market considering renewable

generation uncertainty," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 798–809, Jan. 2021.
[28] J. Kang, R. Yu, X. Huang, S. Maharjan, Y. Zhang, and E. Hossain, "Enabling localized peer-to-peer electricity trading among plug-in hybrid electric vehicles using consortium blockchains," *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 3154–3164, Dec. 2017.
[29] L. Thomas, Y. Zhou, C. Long, J. Wu, and N. Jenkins, "A general form of smart contract for decentralized energy systems management," *Nature Energy*, vol. 4, no. 2, pp. 140–149, Jan. 2019.
[30] M. T. Devine and P. Cuffe, "Blockchain electricity trading under demurrage," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 2323–2325, Mar. 2019.
[31] C. A. Floudas and X. Lin, "Mixed integer linear programming in process scheduling: Modeling, algorithms, and applications," *Ann. Oper. Res.*, vol. 139, no. 1, pp. 131–162, Oct. 2005.
[32] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2010, vol. 3, no. 1.
[33] T. Chen and S. Bu, "Realistic Peer-to-Peer Energy Trading Model for Microgrids using Deep Reinforcement Learning," in *Proc. IEEE PES Innov. Smart Grid Technol. Eur. (ISGT-Europe'2019)*, Bucharest, Romania, pp. 1–5.
[34] T. Chen and W. Su, "Indirect Customer-to-Customer Energy Trading with Reinforcement Learning," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 4338–4348, Jul. 2019.
[35] X. Lu, X. Xiao, L. Xiao, C. Dai, M. Peng, and H. V. Poor, "Reinforcement Learning-Based Microgrid Energy Trading With a Reduced Power Plant Schedule," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10 728–10 737, Dec. 2019.
[36] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR'2016)*, San Juan, Puerto Rico, pp. 1–14.
[37] S. Thrun and A. Schwartz, "Issues in Using Function Approximation for Reinforcement Learning," in *Proc. 4th Connect. Model. Summer. 1993*, Hillsdale, New Jersey, pp. 255–263.
[38] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS'2017)*, Long Beach, CA, USA, pp. 6379–6390.
[39] S. Fujimoto, H. V. Hoof, and D. Meger, "Addressing Function Approximation Error in Actor-Critic Methods," in *Proc. 35th Int. Conf. Mach. Learn. (ICML'2018)*, Stockholm, Sweden, pp. 2587–2601.
[40] R. Khurmi and R. Sedha, *Materials Science*, 5th ed. S. Chand & Company Ltd, 2014.
[41] G. Glenk and S. Reichelstein, "Economics of converting renewable power to hydrogen," *Nat. Energy*, vol. 4, no. 3, pp. 216–222, Feb. 2019.
[42] R. Dixon, J. Li, and M. Wang, "Progress in hydrogen energy infrastructure development—addressing technical and institutional barriers," in *Compendium of Hydrogen Energy*. Elsevier, 2016, pp. 323–343.
[43] GOV.UK. (2021) £166 million cash injection for green technology and 60,000 UK jobs. [Online]. Available: https://www.gov.uk/government/news/166-million-cash-injection-for-green-technology-and-60000-uk-jobs
[44] Energy Digital. (2020) Ofgem approves SGN hydrogen homes project in Scotland. [Online]. Available: https://energydigital.com/oil-and-gas/ofgem-approves-sgn-hydrogen-homes-project-scotland
[45] A. Züttel, "Hydrogen storage methods," *Naturwissenschaften*, vol. 91, no. 4, pp. 157–172, Mar. 2004.
[46] S. M. Saba, M. Müller, M. Robinius, and D. Stolten, "The investment costs of electrolysis–a comparison of cost studies from the past 30 years," *Int. J. Hydrog. Energy*, vol. 43, no. 3, pp. 1209–1223, 2018.
[47] M. Schalenbach, G. Tjarks, M. Carmo, W. Lueke, M. Mueller, and D. Stolten, "Acidic or alkaline? towards a new perspective on the efficiency of water electrolysis," *J. Electrochem. Soc.*, vol. 163, no. 11, p. F3197, Aug. 2016.
[48] M. L. Stein, *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.
[49] H. Demsetz, "The cost of transacting," *Q. J. Econ.*, vol. 82, no. 1, pp. 33–53, Feb. 1968.
[50] W. Tushar, T. K. Saha, C. Yuen, D. Smith, and H. V. Poor, "Peer-to-Peer Trading in Electricity Networks: An Overview," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3185–3200, Jul. 2020.
[51] Y. Wu, X. Tan, L. Qian, D. H. K. Tsang, W.-Z. Song, and L. Yu, "Optimal pricing and energy scheduling for hybrid energy trading market in future smart grid," *IEEE Trans Industr Inform*, vol. 11, no. 6, pp. 1585–1596, Dec. 2015.
[52] M. R. Alam, M. St-Hilaire, and T. Kunz, "Peer-to-peer energy trading among smart homes," *Appl. Energy*, vol. 238, pp. 1434–1443, Mar. 2019.
[53] W. Tushar, B. Chai, C. Yuen, S. Huang, D. B. Smith, H. V. Poor, and Z. Yang, "Energy storage sharing in smart grid: A modified auction-based approach," *IEEE Trans. Smart Grid*, vol. 7, no. 3, pp. 1462–1475, May 2016.
[54] S. Sharma, A. Verma, Y. Xu, and B. K. Panigrahi, "Robustly Coordinated Bi-level Energy Management of a Multi-Energy Building under Multiple Uncertainties," *IEEE Trans. Sustain. Energy*, vol. 12, no. 1, pp. 3–13, Jan. 2021.
[55] H. Li, Z. Wan, and H. He, "Constrained EV Charging Scheduling Based on Safe Deep Reinforcement Learning," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2427–2439, May 2020.
[56] Y. Jiang, C. Wan, C. Chen, M. Shahidehpour, and Y. Song, "A Hybrid Stochastic-Interval Operation Strategy for Multi-Energy Microgrids," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 440–456, Jan. 2020.
[57] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–33, Feb. 2015.
[58] H. V. Hasselt, "Double Q-learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS'2010)*, Vancouver, Canada, pp. 2613–2621.
[59] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.
[60] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel, "Benchmarking deep reinforcement learning for continuous control," in *Proc. 33th Int. Conf. Mach. Learn. (ICML'2016)*, New York City, NY, USA, pp. 1329–1338.
[61] Pecan Street Inc. (2018) Dataport. [Online]. Available: https://www.pecanstreet.org/dataport/
[62] Office of Energy Efficiency & Renewable Energy. (2011) Commercial and Residential Hourly Load Profiles for all TMY3 Locations in the United States. [Online]. Available: https://openei.org/datasets/dataset/
[63] RWTH Aachen University. (2014) Smart Energy Data: Aachen/ Cologne Virtual Power Plant. [Online]. Available: https://data.lab.fiware.org/organization/rwth-aachen-university
[64] ISO New England. (2018) Electricity price. [Online]. Available: https://www.iso-ne.com/
[65] U.S. Energy Information Administration. (2018) United States Natural Gas Industrial Price. [Online]. Available: https://www.eia.gov/dnav/ng/hist/n3035us3m.htm
[66] ——. (2018) U.S. Energy-Related Carbon Dioxide Emissions. [Online]. Available: https://www.eia.gov/environment/emissions/carbon/
[67] H. R. Ellamla, I. Staffell, P. Bujlo, B. G. Pollet, and S. Pasupathi, "Current status of fuel cell based combined heat and power systems for residential sector," *J. Power Sources*, vol. 293, pp. 312–328, Oct. 2015.

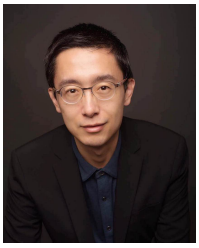


Tianyi Chen (Student Member, IEEE) received the M.Sc. degree from University of Southampton, Southampton, U.K., in 2017. He is currently pursuing a Ph.D. degree with University of Glasgow, UK. His current research interests include the P2P energy trading, multi-energy system, machine learning optimisation and deep reinforcement learning application.

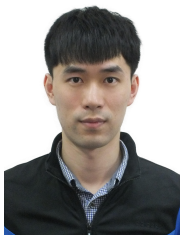
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3124465, IEEE Transactions on Smart Grid

Shengrong Bu received her Ph.D. degree in Electrical and Computer Engineering from Carleton University, Canada, Masters by Research degree in Electrical Engineering from University of Wollongong, Australia, and B.E. in Mechanical Engineering and Automation from Huazhong University of Science and Technology, China.

From 2012 to 2014, she held a research position at Huawei Technologies Canada Inc., Ottawa, as an NSERC Industrial R&D Fellow. From 2014 to 2021, she was a Lecturer (Assistant Professor) with the James Watt School of Engineering, University of Glasgow, UK. Currently, she is an Associate Professor at Brock University, Canada. Her research interests include multi-vector energy microgrids, smart grids, future wireless networks, wireless network security, big data analytics, deep reinforcement learning, and game theory. She was a recipient of three best paper awards at IEEE international conferences. Highlights of her professional activities include duties as a peer reviewer for EPSRC and Carnegie Trust, an Associate Editor for Wireless Networks (Springer) and a Topic Editor for Energies, the TPC Co-Chair for six international workshops or conference symposiums, and duties as the N2Women Mentoring Co-Chair.



Xue Liu is a professor in the School of Computer Science and a William Dawson Scholar at McGill University. He obtained his Ph.D. in Computer Science from The University of Illinois at Urbana Champaign and his B.S. degree in Mathematics and M.S. degree in Automatic Control both from Tsinghua University, China. He has published over 250 research papers in major highly reputable international academic journals and conference proceedings. He has also authored 3 books/monographs and several book chapters. Dr. Liu has served on the organizing committees of 40 major international conferences and workshops. Dr. Liu is an editor/associate editor of IEEE/ACM Transactions on Networking (ToN), ACM Transactions on Cyber-Physical Systems (TCPS), IEEE Transactions on Vehicular Technology (TVT), and IEEE Communications Surveys and Tutorial (COMST). Dr. Liu is a member of the Tau Beta Pi Engineering Honorary Society, a lifetime member of the ACM, a member of the IEEE, and USENIX.



Jikun Kang received the B.Sc degree from Kunming University of Science and Technology, Kunming, China, in 2014, the M.A.Sc degree from Northeastern University, Shenyang, China in 2017, and he is currently pursuing the Ph.D. degree in McGill University, Montreal, Canada. His research interests include reinforcement learning, meta-reinforcement learning, and applied machine learning.

F. Richard Yu (Fellow, IEEE) received the PhD degree in electrical engineering from the University of British Columbia (UBC) in 2003. From 2002 to 2006, he was with Ericsson (in Lund, Sweden) and a start-up in California, USA. He joined Carleton University in 2007, where he is currently a Professor. He received the IEEE TCGCC Best Journal Paper Award in 2019, Distinguished Service Awards in 2019 and 2016, Outstanding Leadership Award in 2013, Carleton Research Achievement Awards in 2012 and 2021, the Ontario Early Researcher Award (formerly Premiers Research Excellence Award) in 2011, the Excellent Contribution Award at IEEE/IFIP TrustCom'10, the Leadership Opportunity Fund Award from Canada Foundation of Innovation in 2009 and the Best Paper Awards at IEEE ICNC'18, VTC'17 Spring, ICC'14, Globecom'12, IEEE/IFIP TrustCom'09 and Int'l Conference on Networking'05. His research interests include connected/autonomous vehicles, security, artificial intelligence, blockchain and wireless cyber-physical systems.

He serves on the editorial boards of several journals, including Co-Editor-in-Chief for Ad Hoc & Sensor Wireless Networks, Lead Series Editor for IEEE Transactions on Vehicular Technology, IEEE Communications Surveys & Tutorials, and IEEE Transactions on Green Communications and Networking. He has served as the Technical Program Committee (TPC) Co-Chair of numerous conferences. He has been named in the Clarivate Analytics list of "Highly Cited Researchers" in 2019 and 2020. He is an IEEE Distinguished Lecturer of both Vehicular Technology Society (VTS) and Comm. Society. He is an elected member of the Board of Governors of the IEEE VTS and Editor-in-Chief for IEEE VTS Mobile World newsletter. Dr. Yu is a registered Professional Engineer in the province of Ontario, Canada, and a Fellow of the IEEE, Canadian Academy of Engineering (CAE), Engineering Institute of Canada (EIC), and Institution of Engineering and Technology (IET).

Zhu Han (S'01–M'04-SM'09-F'14) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively.

From 2000 to 2002, he was an R&D Engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is a John and Rebecca Moores Professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston, Texas. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. Dr. Han received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, IEEE Leonard G. Abraham Prize in the field of Communications Systems (best paper award in IEEE JSAC) in 2016, and several best paper awards in IEEE conferences. Dr. Han was an IEEE Communications Society Distinguished Lecturer from 2015-2018, AAAS fellow since 2019 and ACM distinguished Member since 2019. Dr. Han is 1% highly cited researcher since 2017 according to Web of Science. Dr. Han is also the winner of 2021 IEEE Kiyo Tomiyasu Award, for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: "for contributions to game theory and distributed management of autonomous communication networks."