Data-driven Dynamical Control for Bottom-up Energy Internet System

Haochen Hua, *Member, IEEE,* Zhaoming Qin, Nanqing Dong, Yuchao Qin, Maojiao Ye, *Member, IEEE,* Zidong Wang, *Fellow, IEEE,*, Xingying Chen, *Senior Member, IEEE,* and Junwei Cao, *Senior Member, IEEE,*

***Abstract*—With the increasing concern on climate change and global warming, the reduction of carbon emission becomes an important topic in many aspects of human society. The development of energy Internet (EI) makes it possible to achieve better utilization of distributed renewable energy sources with the power sharing functionality introduced by energy routers (ERs). In this paper, a bottom-up EI architecture is designed, and a novel data-driven dynamical control strategy is proposed. Intelligent controllers augmented by deep reinforcement learning (DRL) techniques are adopted for the operation of each microgrid independently in the bottom layer. Moreover, the concept of curriculum learning (CL) is integrated into DRL to improve the sample efficiency and accelerate the training process. Based on the power exchange plan determined in the bottom layer, considering the stochastic nature of electricity price in the future power market, the optimal power dispatching scheme in the upper layer is decided via model predictive control. The simulation has shown that, under the bottom-up architecture, compared with the conventional methods such as proportional integral and optimal power flow, the proposed method reduces overall generation cost by 7.1% and 37%, respectively. Meanwhile, the introduced CL-based training strategy can significantly speed up the convergence during the training of DRL. Last but not least, our method increases the profit of energy trading between ERs and the main grid.**

***Index Terms*—Bottom-up, deep reinforcement learning, energy Internet, microgrid, stochastic system.**

I. Introduction

The energy Internet (EI) aims to introduce some of the advanced features of the Internet to the energy system, such that any legitimate subject can freely obtain access to the system and share information and energy with other subjects [1]. Open and peer-to-peer energy supply and sharing are the main features of the EI. The integration of both information and energy is the key to achieving an information-led and highly controllable energy system [2], [3]. Through the combination of the Internet and renewable energy sources (RESs), the EI converts the centralized supply of energy into a distributed supply of energy; that is, each local area can make full use of its own solar power, wind power, natural gas, etc., such that the energy supply in each area is relatively independent.With the hope of acquiring an integration of distributed, intermittent and diversified energy supply and demand, this approach aims to construct a type of energy network in which RESs are efficiently utilized and the increasing energy demand is met, whereas the damage to the environment in the process of energy utilization is reduced [4].

In many practical EI scenarios, MGs are designed to operate jointly in the sense that information and energy are exchanged via a new type of electrical device called ERs (also called energy hubs, or electric routers) [5], [6]. It is pointed out that ERs are core devices in the future power and energy system as the routers in the Internet [7], [8]. As an important direction for the future development of energy service industry, integrated energy service will be the key means to implement supply side structural reform and promote demand side response [9]. In order to guarantee reliable integrated energy service, the energy management related research in the field of EI has attracted much attention, resulting in significant advances in the past five years; see, e.g., [10]–[13]. In particular, when the investigated EI scenario is composed of interconnected MGs in which power generation mainly originates from RESs, the system's energy management issue is much more challenging than that of the conventional utility grids [14], [15].

The conventional energy management scheme follows a top-down mode, which requires a centralized controller to decide the operation of each component within the power system [16]. Though such characteristic allows higher efficiency for the utilization of various energy sources, it requires the associated global optimization problems to be accurately solved with limited availability of time and computing resources. With the development of edge computing technologies, more intelligent devices have been introduced into existing energy systems, which further contributes to the complexity of energy management problems in EI scenarios. For the system composed of multiple MGs, as the number of MGs increases, the centralized control scheme would suffer from a huge computation burden, due to the high-dimensional searching space of potential operation strategies [17], [18]. In the meantime, the top-down architecture may not be efficient enough when dealing with

This work is supported in part by the National Natural Science Foundation of China under Grant No. 52107089, in part by the Fundamental Research Funds for the Central Universities of China under Grant No. B200201071, in part by the National Natural Science Foundation of China under Grant No. 62173181, in part by the Fundamental Research Funds for the Central Universities under Grant No. 30920032203, and in part by the BNRist Program under Grant No. BNR2021TD01009.

H. Hua and X. Chen are with the College of Energy and Electrical Engineering, Hohai University, Nanjing, 211100, China; Z. Qin is with the Department of Automation, Tsinghua University, Beijing, China; N. Dong is with the Department of Computer Science, University of Oxford, Oxford, OX1 3QD, UK. Y. Qin is with the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, CB3 0WA, UK; M. Ye is with the School of Automation, Nanjing University of Science and Technology, 210094, China; Z. Wang is with the Department of Computer Science, Brunel University London, Uxbridge, UB8 3PH, UK; J. Cao are with Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, 100084, China;.

Haochen Hua and Zhaoming Qin are co-first authors. Corresponding author: Junwei Cao, email: jcao@tsinghua.edu.cn.

a power system with high proportion of distributed energy resources, especially for the EI system powered by millions of intelligent power electronic devices [18]. In the conventional top-down mode, customers' role is subject to the upper control strategy, and the personalized needs are ignored [19]. The centralized controller issues instructions to control all the energy interaction processes, and no free transaction can be directly made between end users. Therefore, flexible energy transactions between consumers in the power market are limited. For instance, considering various customized demands from local agents and heterogeneous power consumption behavior, it could be difficult to find a proper centralized controller to meet the desired target. Thus, a bottom-up control scheme would be a more feasible option, and it would benefit the privacy protection for customers as well [3], [18].

Via the coordinative operation of intelligent agents at the MG level and ER level, energy management solutions regarding the future power and energy system generally come along with a multi-layer (or, hierarchical) architecture; see, e.g., [15], [16], [20]–[25]. In [15] a distributed control scheme for multiple MGs is designed to manage the balance between control performance and computation feasibility. Jing et al. [16] present a problem-oriented hierarchical approach, such that the operation cost of urban energy system is reduced by 15%. Jain et al. [20] study a data-driven method for distributed energy resources and reduce the levelized cost of electricity to nearly 50%. Under a layered architecture, energy system control problems are solved more efficiently as well as effectively compared with the ones without being layered. In [21], it is pointed out that a bottom-up layered energy management strategy is urged to be implemented in future EI scenarios. Yi et al. [22] have investigated the bottom-up optimization model for the inter-regional power grid planning in China. Besides, a class of bottom-up optimization models have been used to plan for the Greek power supply [23].

Although the aforementioned studies have analysed the hierarchical control scheme and bottom-up model, the research focus of these works are different from that of this paper, and there exist some defects among these works. To illustrate, the perspective of building a bottom-up energy infrastructure has been reported in [18] without providing theoretical analysis or numerical solution. In [21] a stochastic control strategy has been studied to achieve the bottom-up management mode for one typical MG only. The main drawback of [21] is that energy operators are not provided with an overall systematic control scheme for the whole EI system. Besides, the dynamical programming approach adopted in [21] cannot be directly applied to relatively large dimensional control systems. The authors in [22] and [23] mainly focus on the overall long-term planning for power grid and energy system, lacking dynamical control for short-term system analysis. Although the principle of bottom-up energy management has been considered in [26], the system architecture is designed without being layered, which is relatively restrictive.

In recent years, with the development of advanced metering infrastructures, available data volume has been becoming higher than ever before, which benefits designing and management of the future energy systems [27]. Accompany with the progress of modern artificial intelligence (AI) technology, the data-driven energy management scheme has started to show its superiority to the conventional mathematical model-based method. For example, when the neural networks in combination with stochastic analysis is used to model the predicted power of photovoltaic panels (PVs) and loads, the energy control result with respect to the regional EI has been shown to be better than that with the conventional deterministic methods [21]. In [26], a so-called model-free approach has been implemented, and by exploiting an asynchronous advantage actor-critic (A3C) algorithm, better control effects have been achieved compared with the conventional optimal power flow (OPF) method. In [28], a model-free method, which is data-driven, has been adopted to voltage control, such that system stability can be realized. In [29], based on deep neural network techniques, an intelligent multi-MG energy management method is applied to increase the profitability of electricity trading. Furthermore, the deep reinforcement learning (DRL) method has now been regarded as a powerful tool to solve such data-driven optimization issues, and reader can refer to [30], [31] and the references therein for more examples. Compared with the traditional multi-stage optimization methods, DRL has unparalleled advantages in end-to-end training, utilizing global information and continuous control [32]. In this work, we leverage the advances in DRL to implement the proposed bottom-up system control idea. It is notable that the studied bottom-up energy management problem in this paper has *not* been studied via data-driven methods.

The large amount of available data, the progress of AI technology, and the requirement to establish bottom-up management rules for EI have prompted us to carry out the research in this paper. It is worth mentioning that the flexible energy trading between customers in the electricity market is essential for the future EI [4], which ought to be fully considered in the power dispatching scheme. In order to address the above issues, in this article, a two-tier EI architecture is proposed. In such multi-layer structure, the closely related MGs are considered as a small system, namely MG clusters. With the purpose of total cost reduction and power balance assurance, in the bottom layer, MGs in the MG clusters are designed to operate coordinately and share energy via ER networks within clusters. Noticing that the power/energy link between MGs and ERs is bi-directional, MGs are thus capable of exchanging energy with ERs when necessary. Meanwhile, in the upper layer, the optimal power dispatching scheme considering the transmission loss in the ER network is calculated with the energy exchange requests provided by the bottom layer.

For such a bottom-up system management issue, the solution is obtained via two main steps. Firstly, in the bottom layer, each MG in the considered MG cluster is controlled independently by minimizing its own objective function via DRL approach. The power exchange plans between these MGs and the corresponding ERs are then determined. Despite the advances of DRL, it suffers the sample efficiency challenge in the field of power system: the interactions with the power system are extremely expensive while DRL relies on a massive interactions. To tackle this issue, we incorporate the concept of *curriculum learning* (CL) [33] into DRL. It is worth mention-

ing that the proposed CL-based training strategy is not specific to a particular DRL method as it could be easily extended to most existing DRL methods. Next, in the upper layer, the optimal power dispatching scheme is decided regarding the power exchange plans in the bottom layer and the electricity price in the energy market. The stochastic analysis and model predictive control (MPC) techniques are used to solve the upper layer energy control problem. The main contribution and importance of this work is outlined as follows.

1) Focusing on a regional EI scenario, this is the very first time that dynamical control strategies are designed to realize a data-driven bottom-up energy management scheme. Thereby, the research outputs of this paper can be viewed as a supplement to the existing literatures focusing on bottom-up energy management from different perspectives; see, e.g., [16], [21]–[23], [26]. By adopting such a bottom-up system control strategy, not only better performance regarding minimizing the overall operation cost can be achieved, but also the role of customers can be effectively transformed from passive to active.
2) The adopted data-driven AI technique, which refers to the DRL method with the advantage actor-critic (A2C) algorithm [34], effectively avoids the system modeling errors and the parameter estimation deviations. In contrast to the conventional complex modeling process with respect to RESs and loads, the proposed method can skip the dynamical modeling process and therefore is regarded as an advantage over the existing works [15], [16], [21], [25], [34]. When solving the considered high-dimensional stochastic energy management problem, we demonstrate that the adopted DRL algorithm is more advantageous than the conventional approaches such as proportional integral (PI) based and OPF based control schemes.
3) To facilitate the training process of DRL, we propose a training strategy of DRL based on CL. The networks for a single MG are trained independently to learn an optimal policy for this MG. Then, the learned policy is transferred to other MGs as the initial policies for the corresponding training. Note, due to the similarities of the environment between MGs, these transferred policies are already near-optimal for the other MGs. With such a novel design, the backbone DRL model converges much faster than the vanilla DRL model, and the cost of training, including the expensive interaction with the environment, is largely reduced. We want to emphasize that, instead of improving the convergence speed at the expense of control effects, the empirical results show that the proposed training strategy can also improve the learning performance. To the best of our knowledge, this is is first study of incorporating CL with DRL in the field of power and energy system.
4) The stochastic process driven by geometric Brownian motion is used to model the electricity price, allowing for massive stochastic energy transactions to be implemented in the proposed bottom-up EI framework, which is innovative as well as instructive in building a flexible electricity market for the future EI. In this sense, the power dispatching scheme proposed in this paper demonstrates an effective way to achieve better energy sharing performance in an electricity market with stochastic price deviations.

The remainder of this paper is organized as follows. In Section II, we present the system modeling. Then, the bottom-up management issue is formulated in Section III. In Section IV, we introduce the proposed CL-based DRL and MPC approach. Finally, in Section V and Section VI, we provide the numerical results and draw the conclusions, respectively.

II. System Description

A. System Architecture

Following the EI paradigm, in this paper, the operation of a typical MG cluster in the EI system is investigated. As shown in Fig. 1, there exist multiple MGs in the cluster where each MG is connected to an ER which enables the flexible energy exchange within the cluster. Additionally, the ER network inside the given cluster is interconnected with the main grid. At time $t \in [0, T]$, each MG in the bottom layer determines the amount of the exchanged power based on its own states. Next, the ERs in the upper layer receive the power exchange information and decide the power dispatching between ERs and the main grid.

The ER network structure of the considered MG cluster can be captured by an undirected graph denoted as $\mathcal{G}(\mathcal{V}, \xi)$. ERs are represented by the vertex set $\mathcal{V} = \{1, 2, \ldots, n\}$, and transmission lines between ERs are denoted as edges in $\xi \subseteq \mathcal{V} \times \mathcal{V}$. n denotes the number of MGs. Without loss of generality, each MG is assumed to be composed of a subset of the following components, PVs, micro-turbines (MTs), fuel cells (FCs), diesel engine generators (DEGs), battery energy storage devices (BESs), as well as local loads. Accordingly, let us denote the i-th MG in the considered MG cluster as MG_i. In addition, multiple bi-directional connections to the utility grid are comprised in such graph representation as well.

B. Bottom Layer

In this paper, power output of PVs and power demand of loads are considered to be uncontrollable. Due to the stochastic and uncertain nature of their power dynamics, it is impractical to precisely model or forecast the power deviations by deterministic models. To address this problem, real-world power data records of PVs and loads in [35] are utilized in our proposed solution to the operation of MG clusters, which could greatly improve the reliability of the studied enengy management method. More specifically, the dynamical models for the power of PVs and loads introduced in [21] are adopted in this paper as follows. For the power of PVs,

$$P^{PV}(t) = \hat{P}^{PV}(t) + \tilde{\alpha} P^{Solar}(t) r^{PV}(t), \tag{1}$$

where $P^{PV}(t)$ refers to the modeled power generation of PVs at time t; $\hat{P}^{PV}(t)$ denotes the power forecasting for PVs at time t; $P^{Solar}(t)$ denotes the solar irradiation at time t; $\tilde{\alpha}$ is a coefficient related to the configuration of PVs in a given

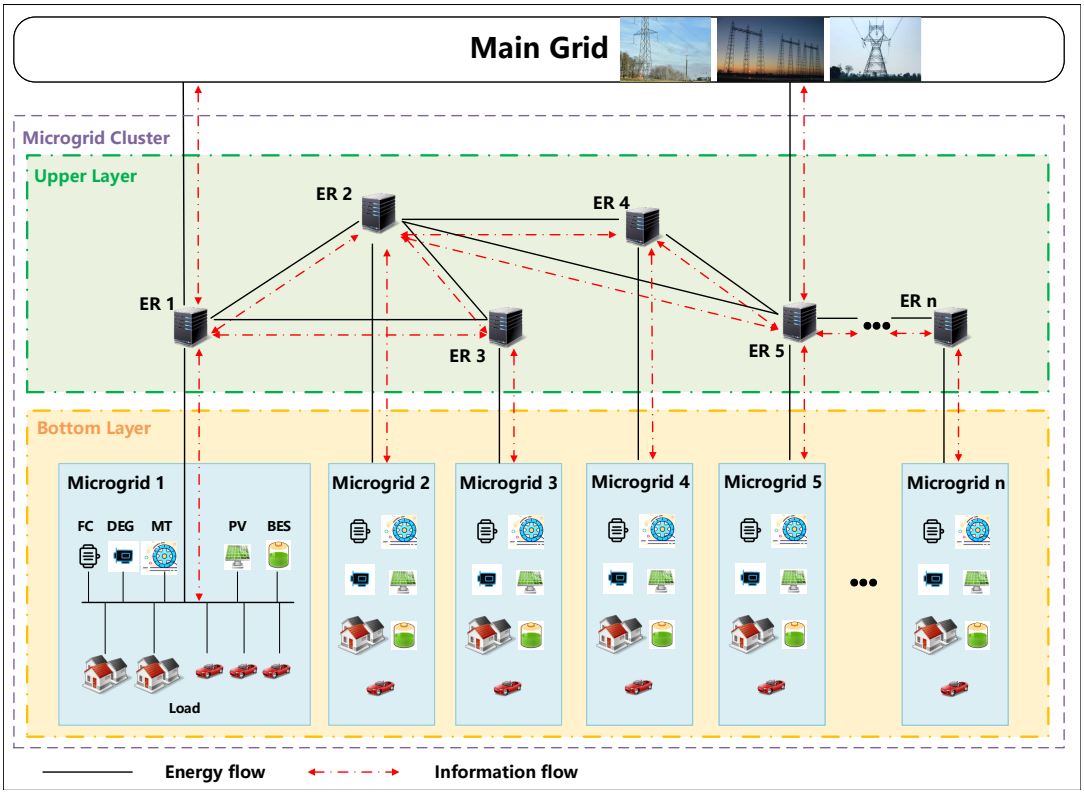


Fig. 1: The considered MG cluster.

MG; and $r^{PV}(t)$ captures the randomness in the PV power generation via an Ornstein-Uhlenbeck process [36]. For the detailed procedure of obtaining dynamical model (1) via real data, readers can refer to [21].

Similarly, the power model of loads via real data can be represented as $P^L(t) = \hat{P}^L(t) + \tilde{\beta}e^L(t)$, where $P^L(t)$ is the modeled power output of loads; $\hat{P}^L(t)$ is the overall trend of load power; $\tilde{\beta}$ is a constant related to the magnitude of the load power deviations; $e^L(t)$ is driven by Ornstein-Uhlenbeck process, which represents stochastic deviations existing in the load power.

Controllable distributed generators (DGs), i.e., MTs, FCs, DEGs, are generally applied for the power balance maintenance in the system under consideration. Hence, controllers are set in controllable DGs to adjust their power output accordingly. Given $g \in \{\text{MT, FC, DEG}\}$, let us denote $P_i^g(t)$ as the power output of controllable DGs in the i-th MG of the considered MG cluster, and denote $u_i^g(t)$ as the corresponding control signal. The power dynamics of DGs are formulated as ordinary differential equations [37],

$$dP_i^g(t) = -\frac{1}{T_i^g}(P_i^g(t) - B_i^g u_i^g(t))dt, \tag{2}$$

where T_i^g stands for the time constant of controllable DGs of the i-th MG; B_i^g is the maximum allowed value of $P_i^g(t)$.

For the i-th MG of the investigated MG cluster, the corresponding charging/discharging power and the state of charge (SOC) for BESs at time t are denoted as $P_i^{BES}(t)$ and $SOC_i(t)$, respectively. According to [38], the dynamics of $SOC_i(t)$ is given in (3),

$$dSOC_i(t) = \frac{\eta_i}{Q_i}P_i^{BES}(t)dt, \tag{3}$$

where Q_i is the capacity of BESs in the i-th MG; $\eta_i(t)$ is the charging/discharging coefficient for BESs and is defined in (4).

$$\eta_i(t) \triangleq \begin{cases} \eta_{0,i} - \eta_{1,i}P_i^{BES}(t)/P_{max,i}^{BES}, & P_i^{BES}(t) \geq 0, \\ 1/\left(\eta_{0,i} + \eta_{1,i}P_i^{BES}(t)/P_{max,i}^{BES}\right), & P_i^{BES}(t) \leq 0. \end{cases} \tag{4}$$

The coefficients $\eta_{0,i}$ and $\eta_{1,i}$ in (4) are related to the characteristics of BESs; see, e.g., [38]. In many practical scenarios, the SOC is restricted within an appropriate range for better reliability. As a protection for BESs from being damaged by large energy throughput, the input/output power of BESs shall be constrained as well. In this paper, the constraints for $P_i^{BES}(t)$ and $SOC_i(t)$ are set as $-P_{max,i}^{BES} \leq P_i^{BES}(t) \leq P_{max,i}^{BES}$ and $SOC_i^{min} \leq SOC_i(t) \leq SOC_i^{max}$, where $P_{max,i}^{BES}$, SOC_i^{min} and SOC_i^{max} denote the maximum allowed absolute value of P_i^{BES}, minimum and maximum allowed value of SOC_i, respectively.

On top of the models introduced above, the power balance in MG_i is considered. Let us denote $P_i^L(t)$ and $P_i^{PV}(t)$ as the demand of loads and output power of PVs in MG_i at time t, respectively. When BESs and power exchange with corresponding ER is not considered, $\Delta P_i(t)$, the power mismatch in MG_i at time t, is expressed as

$$\Delta P_i = P_i^{PV} + P_i^{MT} + P_i^{FC} + P_i^{DEG} - P_i^L. \tag{5}$$

For notation simplicity, time t is omitted in the above equation and most of the formulas thereafter.

To maintain the power balance in each MG, we firstly set P_i^{BES} as the power mismatch ΔP_i. Once P_i^{BES} violates its constraints, it will then be set as the maximum or minimum allowed value accordingly. Similar rules are applied for the cases when constraints for SOC are violated. Secondly, the power exchange P_i^{exc} between MG_i and ER_i is introduced to guarantee power balance in the considered scenario with $P_i^{exc} = P_i^{BES} - \Delta P_i$. Note that the power exchange with ER exists only when the power of BESs reaches the power limit, which embodies the principle of local power balance shall be autonomously achieved with priority [3], [21]. Finally, the power balance of MG_i is expressed as follows,

$$P_i^{exc} + P_i^{PV} + P_i^{MT} + P_i^{FC} + P_i^{DEG} = P_i^{L} + P_i^{BES}. \quad (6)$$

C. Upper Layer

In the upper layer, the power exchange among MGs inside the considered MG cluster is achieved via the ER network. Meanwhile, the MG cluster is assumed to be able to trade energy with the utility grid through ERs. The similar scenario regarding energy trading via ER has been considered in [39] and the references therein. Thus, power can be transmitted from MGs with surplus energy to those lacking of energy. In some extreme situations, the ER network may purchase or sell energy via connections with the utility grid to assure the stable operation of the MG cluster.

For the power dispatching in the upper layer, the power flow from ER_i to ER_j is denoted by $P_{i,j}^{ER}$. A negative value of $P_{i,j}^{ER}$ means the power is transmitted from ER_j to ER_i, and vice versa. Thus, $P_{i,j}^{ER} = -P_{j,i}^{ER}$ holds naturally. We denote the set of indexes for ERs connected to the utility power grid as $L = \{l_1, l_2, \ldots, l_m\}$, where m is the total number of ERs connected to the utility grid. Considering the power balance at the node ER_i, we have

$$P_i^{exc} + \sum\nolimits_{(i,j)\in\xi} P_{i,j}^{ER} + P_i^{m} = 0, \quad \forall i \in V, \quad (7)$$

where P_i^m represents the power exchange between ER_i and the main gird. Typically, we have $P_i^m = 0, i \notin L, \forall i \in V$ for the ERs that are not directly connected with the utility power grid. Meanwhile, accounting the capacity of transmission lines, we have the following constraint: $-P_{max,ij}^{ER} \le P_{i,j}^{ER} \le P_{max,ij}^{ER}$, where $P_{max,ij}^{ER} \ge 0$ is the maximum allowed power flow in the transmission line from ER_i to ER_j.

III. Problem Formulation

A. Bottom Layer

In the bottom layer, each MG is considered as an independent entity with its own objectives (i.e., maintenance of local power balance, minimization of overall operation cost). Detailed mathematical formulation for the cost function regarding MG_i during $[0, T]$ is given as follows,

$$J_i = \int_0^T \left[C_i^{MT} + C_i^{FC} + C_i^{DEG} + C_i^{BES}\right] dt, \quad (8)$$

where the operation costs related to controllable DGs and BESs are considered in the integrand, which are further explained in the following.

When the power generation cost of controllable DGs is taken into account, given $g \in \{\text{MT, FC, DEG}\}$, the term C_i^g in (8) refers to the total cost for the power generation of g, and the corresponding cost for each term is formulated in the commonly used quadratic form:

$$C_i^g = a_i^g + b_i^g P_i^g + c_i^g (P_i^g)^2, \quad (9)$$

where a_i^g, b_i^g, c_i^g are coefficients of the corresponding DGs of MG_i, and can be estimated by parameter identification methods. For the purpose of rational utilization of BESs, on the one hand, it is desired that the SOC is maintained around the level

$$SOC_i^{mid} = \frac{1}{2}(SOC_i^{min} + SOC_i^{max}). \quad (10)$$

On the other hand, the input/output power of BESs is expected to be restrained. Therefore, the cost $C_{i,j}^{BES}$ in (8) is formulated as

$$C_i^{BES} = \kappa_{i,1}(SOC_i - SOC_i^{mid})^2 + \kappa_{i,2} {P_i^{BES}}^2, \quad (11)$$

where constants $\kappa_{i,1}$ and $\kappa_{i,2}$ are weight factors.

In this sense, the optimal control problem for the i-th MG can be formulated as

$$\min\nolimits_{u_i^g \in \mathcal{U}} \quad \mathbb{E}\left[J_i\right], \quad \text{s.t.} \quad (2) - (6), \quad (12)$$

where $\mathbb{E}$ denotes the mathematical expectation which is introduced due to the stochastic models, and J_i is the objective function defined as (8). Once problem (12) is solved, the power exchange between the i-th MG and its connected ER can be determined subsequently, and the upper layer would collect the power exchange information for the calculation of energy dispatching scheme in the ER network.

B. Upper Layer

In the upper layer, based on the power exchange between MGs and ERs in the bottom layer of the considered MG cluster, optimal power flows among ERs and the utility power grid are determined thereafter. The key objective is to maximize the profit of power trading and minimize transmission cost for the power dispatching in the considered EI system. The cost function for the upper layer can be formulated as follows,

$$J_{upper} = \int_0^T \left[\sum_{i\in L} C_i^m(t) + \sum_{(i,j)\in\xi} \iota_{ij}(P_{i,j}^{ER}(t))\right] dt, \quad (13)$$

where $C_i^m(t)$ is the cost of electricity exchange between ER_i and the utility power grid at time t; $\iota_{ij}(\cdot)$ is a function representing the transmission loss regarding the ER_i-ER_j link. For simplicity, let us define $\iota_{ij}(P_{i,j}^{ER}) = \mu_{ij} {P_{i,j}^{ER}}^2$, where μ_{ij} denotes the loss coefficient for the ER_i-ER_j link. In (13), $C_i^m(t)$ is defined as

$$C_i^m(t) = -\left[\alpha_i^p(t)\mathbb{I}_{\{P_i^m(t)\le 0\}} + \alpha_i^s(t)\mathbb{I}_{\{P_i^m(t)>0\}}\right] P_i^m(t). \quad (14)$$

Here, the indicator function $\mathbb{I}_{\{x\}}$ is introduced as

$$\mathbb{I}_{\{x\}} \triangleq \begin{cases} 1, & \text{if } x \text{ is } true, \\ 0, & \text{if } x \text{ is } false, \end{cases} \quad (15)$$

where x stands for a logical expression. In (14), $\alpha_i^p(t)$ and $\alpha_i^s(t)$ are the electricity price for energy purchase and sale from the ER network to the utility grid at time t, respectively.

Additionally, temporal deviations of the electricity price are modeled as $\alpha_i^p(t) = \alpha_{i,min}^p + B_i^p(t)$ and $\alpha_i^s(t) = \alpha_{i,min}^s + B_i^s(t)$, where $\alpha_{i,min}^p$ and $\alpha_{i,min}^s$ are the minimum purchase and sale price at ER_i, respectively; $B_i^p(t)$ and $B_i^s(t)$ are geometric Brownian motions modeling the price deviations in the electricity market. It is remarkable that there have been some literatures introducing stochastic processes into electricity pricing and trading; see, e.g., [39], [40].

Moreover, in the upper layer, considering the power balance constraint (7) for each ER, it is required in the ER network that

$$\sum\nolimits_{i=1}^{n} P_i^{exc} + \sum\nolimits_{i\in L} P_i^m = 0 \tag{16}$$

shall be satisfied. In this sense, the power dispatching problem for the upper layer can be expressed as

$$\begin{array}{ll} \min_{P_{i,j}^{ER}(t), P_i^m(t)} & \mathbb{E}[J_{upper}], \\ \text{s.t.} & (7) - (16). \end{array} \tag{17}$$

By solving the stochastic optimal power dispatching problem (17), the operation scheme for the ER network can be obtained.

IV. Solution to the Energy Management Problem

In this section, the energy management problem formulated in Section III is solved via a two-step procedure. Firstly, for the operation of each MG, the optimal control problem (12) is solved with a DRL algorithm. Then, the concept of CL is integrated into DRL to speed up the training of the policies. Finally, the optimal power dispatching scheme for (17) in the upper layer is obtained with convex optimization techniques.

A. DRL for Energy Management of Individual MGs

In our proposed solution, DRL-based controllers are deployed at individual MGs separately. Here, we take MG_i as an example to illustrate the details of the proposed control approach.

The optimal control problem (12) is reformulated as a standard Markov decision process which is widely adopted in reinforcement learning literature. With the time interval Δt, the discretized time steps can be represented as $\mathcal{K} = \{0, 1, \ldots, K\}$, where $K = T/\Delta t$, and T is the terminal time. In this sense, MG_i is viewed as an environment with state

$$\begin{aligned} s = [&k\Delta t, P_i^{PV}(k\Delta t), P_i^L(k\Delta t), P_i^{MT}(k\Delta t), P_i^{FC}(k\Delta t), \\ &P_i^{DEG}(k\Delta t), SOC_i(k\Delta t), P_i^{BES}(k\Delta t)] \in \mathcal{S}, \end{aligned} \tag{18}$$

where $k \in \mathcal{K}$, and $\mathcal{S}$ is the set of all possible system states. Moreover, the action is defined as $u = \left[u_i^{MT}, u_i^{FC}, u_i^{DEG}\right]$. The elements of state s include the time, power of PVs, power of loads, power of controllable DGs, SOC and charge/discharge power of BESs at step k. The transition between the possible states follows a time homogeneous Markov chain, which is a straightforward result from the dynamical model introduced in Section II.

In order to reduce the variance of input for neural network, pre-process for original state is required. Each element of the original state is transformed within the range (0,1). The state of MG after pre-processing $\hat{s}$ is presented as

$$\hat{s} = \left[\frac{k\Delta t}{T}, \frac{P_i^{PV}}{P_{i,max}^{PV}}, \frac{P_i^L}{P_{i,max}^L}, \frac{P_i^{MT}}{B_i^{MT}}, \frac{P_i^{FC}}{B_i^{FC}}, \frac{P_i^{DEG}}{B_i^{DEG}}, SOC_i, \frac{P_i^{BES}}{P_{max,i}^{BES}}\right] \tag{19}$$

where $P_{i,max}^L$ and $P_{i,max}^{PV}$ denote the maximum values of P_i^L and P_i^{PV}, respectively. At time step k, given the full observation $O(s)$ of the system state s (i.e., $O(s) = s$), the DRL algorithm generates the corresponding control input $u \in \mathcal{U}$ from the learned control policy $\pi(u|s)$. The generated controller u is then applied to the environment. Subsequently, the state of the environment s will transit to a new state s' according to its transition probability $P(s'|s, u)$ and produce a reward $r(s, u)$.

Given a policy π, the total accumulated reward from an initial state $s \in \mathcal{S}_i$ is defined as $R^\pi(s) = \sum_{k=0}^{K} \gamma^k r_k$, where $\gamma \in (0, 1]$ is a discounting factor; r_k is the reward from the environment for the state transition from state s_k to state s_{k+1} given action u_k, and we have $s_0 = s$. In this sense, for a policy π, the value function $V^\pi(s)$ is calculated as $V^\pi(s) = \mathbb{E}_\pi[R^\pi(s)], s \in \mathcal{S}$. The DRL algorithm is designed to seek the optimal policy π^* which minimizes the value function $V^\pi(s)$ for all possible $s \in \mathcal{S}$. Thus, with properly selected reward function $r(s, u)$, the value function will be equivalent to the cost function for the i-th MG in (8).

In the meantime, the constraints on the i-th MG are taken into consideration via introducing extra penalty terms in the reward function $r(s, u)$. Considering constraints of BESs, the corresponding penalty function is formulated as

$$\begin{aligned} \phi(s) = \quad & \beta\left[1 - \mathbb{I}_{\{SOC_i^{min} \le SOC_i \le SOC_i^{max}\}}\right] \\ & + (1-\beta) \times \Big[\max(-P_i^{BES} - P_{max,i}^{BES}, 0) \\ & + \max(P_i^{BES} - P_{max,i}^{BES}, 0)\Big], \end{aligned} \tag{20}$$

where $\beta \in (0, 1)$ is a weight factor. In (20), the first part refers to the constraint for SOC, and a fixed penalty will be produced when the constraint for SOC_i is violated. The second part is relevant to the constraint for the power of BESs. This term has linear growth when P_i^{BES} leaves its constrained range, otherwise it will be zero. Noticing that C_i^{MT}, C_i^{FC}, C_i^{DEG} and C_i^{BES} are functions with respect to a part of state s, the reward for state s can be rewritten as

$$\begin{aligned} r(s, u) = & \epsilon_1 \left[C_i^{MT} + C_i^{FC} + C_i^{DEG} + C_i^{BES}\right] \\ & + \epsilon_2 \phi(s), \end{aligned} \tag{21}$$

where coefficients ϵ_1, ϵ_2 need to be properly set for a better performance and faster convergence. For example, when the system is relatively unstable, i.e., the constraints of BESs and SOC are violated, the value of ϵ_2 should be increased appropriately. In this sense, the agents get more penalty when the constraints are violated. Thus, with properly selected coefficients ϵ_1 and ϵ_2, the value function $V^{\pi^*}(s)$ of the optimal policy π^* is just the approximation of the value function of problem (12), and the optimal policy π^* for the formulated Markov decision process would also be an approximated optimal solution for (12).

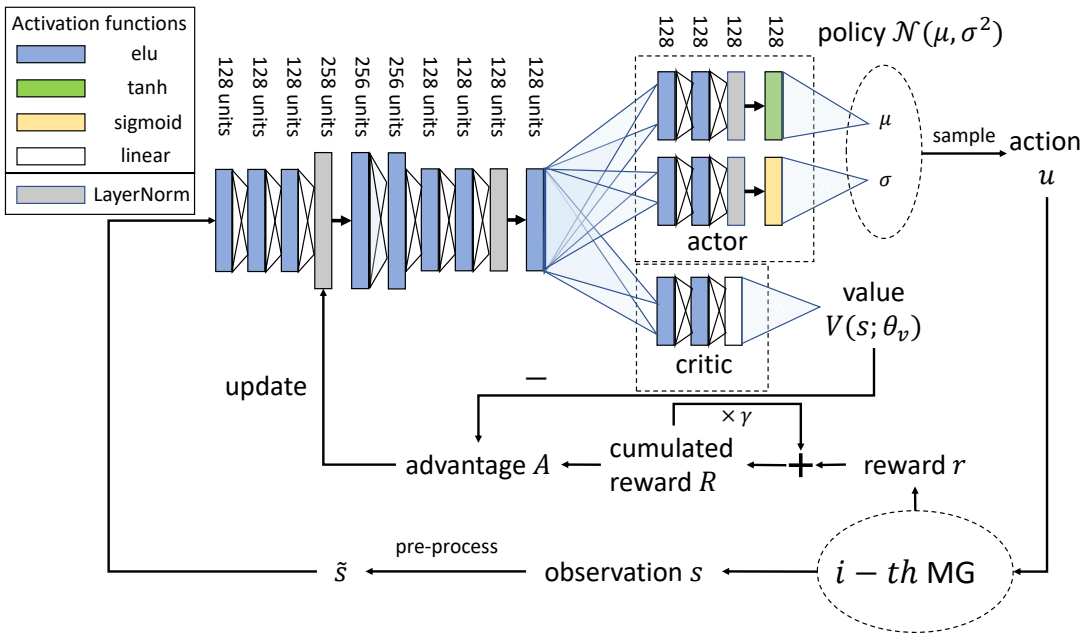


Fig. 2: Scheme of A2C network.

Until now, the optimal control problem (12) has been reformulated under the setting of Markov decision process which can be directly solved via DRL methods. In this paper, the A2C algorithm [34] is adopted, maintaining a policy network (*actor*) parameterized by network parameter θ_π for the policy $\pi(u|s;\theta_\pi)$ and a network (*critic*) with parameter θ_v for the estimation of the value function $V(s;\theta_v)$.

The structure of the actor-critic network adopted in this paper is presented in Fig. 2. In this figure, the observation of the environment is firstly fed to the network after necessary pre-processing procedures. Multi-layer (8-layer) fully connected network is utilized to extract feature vectors from the observation. It is worth mentioning that the depth of network depends on the complexity of the control problem. Usually, more complicated problem requires more complex control strategy which should be approximated by deeper neural network. Then, these features are provided as the input of the *actor* and *critic* networks simultaneously.

Further, with the output $\mu(s;\theta_\pi)$ and $\sigma(s;\theta_\pi)$ of the actor network, the action a is generated based on the policy

$$\pi(u|s;\theta_\pi) = \tanh(\mathcal{N}(\mu(s;\theta_\pi), \sigma(s;\theta_\pi))), \tag{22}$$

where $\tanh(\cdot)$ is the hyperbolic tangent function, and $\mathcal{N}(\mu,\sigma)$ denotes the random variable following a multi-variate normal distribution with mean vector μ and covariance matrix σ. It is notable that we select the activation function of output layer for σ as sigmoid to guarantee the positivity of the covariance matrix.

On the other hand, the output of the critic network is treated as an estimate of the value function $V(s;\theta_v)$. Given outputs of the critic and actor network as well as rewards from interactions with the environment, the loss function for the *critic* and *actor* networks are calculated as

$$L_{value} = \frac{1}{2}\mathbb{E}_{u,s\sim\pi}\left[(R(s,u) - V(s;\theta_v))^2\right], \tag{23}$$

$$L_{policy} = \mathbb{E}_{u,s\sim\pi}\left[A \cdot \log \pi(u|s;\theta_\pi)\right], \tag{24}$$

respectively. Here, A can be viewed as an estimate of the *advantage* over approximated value and is calculated as

$$A = R(s,u) + \gamma V(s';\theta_v) - V(s;\theta_v). \tag{25}$$

In (23), L_{value} measures the difference between the real total reward by sampling and the approximated value by critic network. In (24), L_{policy} stands for the weighted *advantage* under the probability distribution of the current policy, such that the updated network would generate more advantageous policy with a larger probability. Thus, the loss function for the whole network is defined as $Loss = \alpha_{value}L_{value} + \alpha_{policy}L_{policy}$, where, α_{value} and α_{policy} are weight coefficients.

Following Algorithm 1, by using the gradient descent approach, the parameters θ_v and θ_π of the actor-critic network will be continuously improved via interactions between the DRL network and the environment. The advantage function A in (25) can effectively reduce the variance of the overall performance of the trained actor-critic network, which makes it easier for the A2C algorithm to be applied in practice.

B. Curriculum Learning for Deep Reinforcement Learning

The training of separate policies for the energy management of MGs requires a large amount of interactions with each MG, which is computationally expensive. Due to the similarity between the energy management problems of MGs in the bottom layer, the policy learned by one MG can be applied to the energy management of other MGs. In this sense, the concept of CL [33] is integrated in the proposed framework.

Algorithm 1 A2C Training Process for the i-th MG

Input: The environment of the i-th MG; the initial parameters of actor network and critic network, i.e., θ_π and θ_v.
$T_{total} \leftarrow 0$, $k \leftarrow 0$.
repeat
  Get state s_k, $k_{start} \leftarrow k$.
  repeat
    Generate u_k according to $\pi(u_k|s_k;\theta_\pi)$.
    Apply u_k to the environment and receive the reward r_k and next state s_k.
    $k \leftarrow k+1$, $T_{total} \leftarrow T_{total}+1$.
  until $k = K$ or $k - k_{start} = k_{max}$
  $R = \begin{cases} 0, & k = K \\ V(s_k;\theta_v), & k \neq K \end{cases}$
  for $i \in \{k-1, k-2, ..., k_{start}\}$ **do**
    $R \leftarrow r_i + \gamma R$
    Accumulate gradients with respect to θ_π:
    $d\theta_\pi \leftarrow d\theta_\pi + \nabla_{\theta_\pi} \log \pi(u_i|s_i;\theta_\pi)(R - V(s_i;\theta_v))$
    Accumulate gradients with respect to θ_v:
    $d\theta_v \leftarrow d\theta_v + \frac{1}{2}\partial(R - V(s_i;\theta_v))^2/\partial\theta_v$
  Perform update of θ_π using $d\theta_\pi$ and θ_v using $d\theta_v$.
until $T_{total} > T_{max}$
Return θ_π and θ_v.

In RL, the goal of CL is to accelerate the learning of a difficult target task by training on a serious of simpler tasks and transferring the knowledge acquired to the target task [41]. In this work, we use the concept of CL to speed up the training and reduce the computational cost of DRL. Concretely, the energy management problems of MGs can be viewed as a series of similar tasks. The policy learned from the energy management of one MG can be transferred to the energy management of the another MG. We adopt a simple strategy based on the concept of *transfer learning* in the literature of deep learning. That is to say, we first learn the policy for one MG, then transfer the policy learned by the first MG as the initial policy for the other MGs.[1] Specifically, the parameters of actor network and critic network, i.e., θ_π and θ_v, are randomly initialized to train the policy for just one MG. When the first MG is trained, the knowledge of the first MG is transferred to the rest MGs by initializing the parameters of the rest MGs with the parameters of the first MG. The details of the proposed CL-based training strategy is demonstrated in Algorithm 2. The proposed CL-based training strategy can also be understood from a *Bayesian* perspective, i.e., we use the policy learned from a particular MG as a strong prior for the other MGs. Theoretically, the proposed training strategy can significantly improve the sample efficiency of DRL and reduce the training cost with a near-optimal initial policy.

C. Power Dispatching in the Upper Layer

While MGs in the bottom layer operate autonomously at each time step, the power exchange between individual MGs and ERs is determined, and then reported to the upper layer. After receiving the request of power exchange from the individual MGs, the upper layer optimizes the power dispatching scheme (17) with the following Algorithm 3.

First, the random variables including the future purchase prices and selling prices are predicted by calculating the expectations of their distributions. Second, the MPC problem can be built up with the knowledge of models. Then, the MPC problem is solved by the convex optimization tool. Finally, the actions at the first time step are executed.

Note that MPC relies on the knowledge of the power dispatching model in the upper layer, which is much easier to be obtained than the models of MGs.

[1]Note, we illustrate the idea of CL by pre-training only one MG. More MGs could be pre-trained and more complex training procedures could be designed for more robust performance.

Algorithm 2 Curriculum Learning for Energy Management of Multiple MGs

Randomly initialize θ_π and θ_v for the 1st MG.
Execute Algorithm 1 for the 1st MG and save the weights θ_π and θ_v.
$\theta_\pi^1 \leftarrow \theta_\pi$, $\theta_v^1 \leftarrow \theta_v$
for $i = 2$ to n **do**
  Execute Algorithm 1 with initial parameters $\theta_\pi = \theta_\pi^1$ and $\theta_v^1 = \theta_v^1$ and the environment of i-th MG.
  $\theta_\pi^i \leftarrow \theta_\pi$, $\theta_v^i \leftarrow \theta_v$.
Return $\theta_\pi^1, \ldots, \theta_\pi^n$.

Algorithm 3 MPC for Power Dispatching in the Upper Layer

for $t = 0$ to T **do**
  Forecast the future electricity prices $\alpha_i^p(t')$ and $\alpha_i^s(t')$, $t' \in [t, T]$ by calculating their expectations.
  Establish the MPC problem (17) with the prediction for future electricity prices.
  Solve the MPC problem via convex optimization toolbox CVX [42].
  Apply the first step of its optimal control sequence to the power dispatching in the upper layer.

V. Numerical Simulation

It has been illustrated in the literature on smart grid systems that the hierarchical framework has been well suited for the energy management tasks in many practical EI scenarios [3], [16], [18], [20]. Thus, in this paper, we focus on the impacts of specific MG operation schemes on the considered bottom-up structure. To this end, we will first acquire the operation costs in both bottom and upper layers (namely, the overall performance) by using the proposed DRL method and the conventional methods separately. Then, comparison will be made to fully demonstrate the effectiveness and superiority of the proposed bottom-up scheme.

A. Experiment Setup

A MG cluster consisting of eight MGs is considered in the simulation. The parameters for each MG are shown in Table I. Regarding the operation costs in the bottom layer, the

TABLE I: Parameters for Eight MGs

Parameters	**Index of MGs**							
	1	2	3	4	5	6	7	8
T^{MT}(min)	20	21	22	20	19	20	21	20
T^{FC}(min)	25	24	25	26	26	25	24	26
T^{DEG}(min)	30	32	29	31	30	28	30	31
B^{MT}(kW)	200	210	250	190	220	230	200	210
B^{FC}(kW)	300	320	280	310	320	300	310	290
B^{DEG}(kW)	400	380	390	400	420	410	390	400
SOC^{min}	0	0	0	0	0	0	0	0
SOC^{max}	1	1	1	1	1	1	1	1
P^{BES}_{max}(kW)	200	200	200	200	200	200	200	200
Q(kWh)	1000	980	950	1000	1020	1050	1000	950
η_0	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
η_1	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
b^{MT}	0.50	0.52	0.53	0.45	0.47	0.48	0.50	0.51
c^{MT}	0.0050	0.0051	0.0048	0.0049	0.0052	0.0050	0.0051	0.0048
b^{FC}	1.00	1.02	0.98	0.97	1.05	0.95	1.00	1.03
c^{FC}	0.0030	0.0029	0.0029	0.0031	0.0031	0.0030	0.0029	0.0030
b^{DEG}	1.50	1.51	1.48	1.47	1.45	1.54	1.50	1.51
c^{DEG}	0.0020	0.0019	0.0021	0.0019	0.0021	0.0020	0.0019	0.0020
κ_1	100	100	100	100	100	100	100	100
κ_2	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}

TABLE II: Parameters for A2C and PSO.

Parameter	γ	β	α_{value}	α_{policy}	T (min)	ω^{PSO}	c_1^{PSO}	c_2^{PSO}
Value	0.95	0.98	0.1	1	10	0.9	0.5	0.3

performances under the proposed DRL controller, OPF based method [43] and particle swarm optimization (PSO)-based PI controller (introduced in Appendix) in MG_1 are compared. The parameters for A2C and PSO are shown in Table II.

The numerical simulation is performed during the period $[0, 24h]$, and the time resolution for simulation is set to be $\Delta t = 1min$. With the discounting factor γ set to be 0.9, the actor-critic network shown in Fig. 2 is trained by utilizing the power data from the real-world scenario reported in [35]. For the A2C approach, at each time step, features including time step k, power consumption of loads, power generation of PVs and controllable DGs, SOC and charge/discharge power of BESs are fed to the A2C neural network as the input after necessary pre-processing procedures. Actions generated from the operation policy obtained from the DRL network are adopted as control signals for DGs in the considered MG. We divide the 100 days of the historical data into training set and test set, and their proportions are 80% and 20%, respectively.

B. Energy Management of Individual MG

The SOC curves under three methods are shown in Fig. 3. It can be observed that, under the proposed DRL scheme, the BESs are effectively utilized. When the output power of PVs is relatively small during the time period $[0, 9h]$, the BESs are discharged for daily utilization; when the output power of PVs is relatively high during $[9h, 18h]$, the BESs are charged to store energy. In contrast, the SOC determined by OPF based method oscillates frequently, leading to an acceleration of the battery aging process. When the PI controller is applied to the considered scenario, the deviations of SOC curve are constrained within a small range, which illuminates that the function of BESs is not fully utilized.

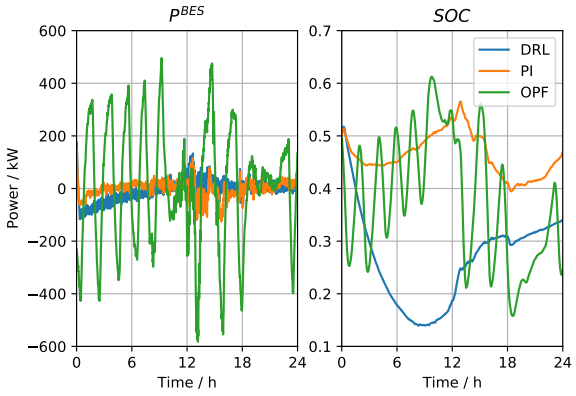


Fig. 3: Power of BESs and SOC curves under three methods.

As shown in Fig. 3, the power input/output of the BESs is effectively maintained in the desired range with relatively smooth deviations by the DRL controller compared with those of the PI controller and OPF method. Focusing on the comparison between DRL and PI controller, it can be observed that the input/output power of BESs under the proposed DRL controller is more stable compared with that under the PI controller, especially during time period $[12h, 18h]$, which shows that DRL controller is more suitable than PI controller to prolong the life of BESs and prevent overuse of BESs.

For the illustrative purpose, the power of DEG under three methods is provided in Fig. 4. It can be observed that the power of DEG under DRL controller is generally lower than

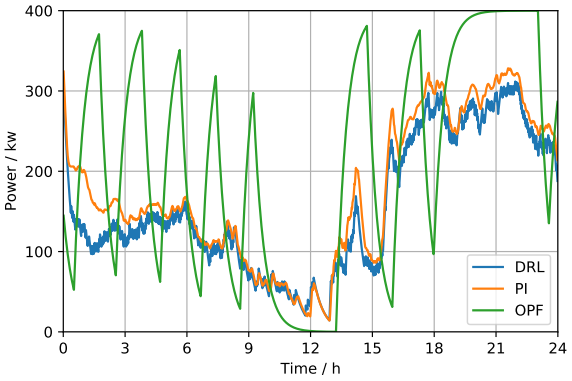


Fig. 4: Power generation of DEG under three methods.

the power under PI method. Especially, during the time period $[13h, 15h]$, the fluctuation of power under DRL controller is relatively small. It is suggested that the proposed DRL controller can reduce the potential risk of over-control, i.e., frequent and drastic adjustment for the power of controllable DGs.

Furthermore, the comparison of the total generation cost in the considered MG shown in Table III illustrates the superior performance of the DRL controller. The total generation cost under the proposed DRL controller is reduced by 7.1% than that under the PI controller, and by 37% than that under the OPF method.

TABLE III: Total Generation Cost of Controllable DGs

Method	Generation Cost
DRL	$(1.354 \pm 0.231) \times 10^4$
PI	$(1.457 \pm 0.245) \times 10^4$
OPF	$(2.150 \pm 0.526) \times 10^4$

C. Effect of Curriculum Learning

To demonstrate the effect of CL in the training of DRL, we depict the training curves of DRL with CL and without CL in Fig. 5. The episode reward is defined as $\sum_{t=0}^{T} r_t$, and the performance during training is evaluated on the test set (20 days of history data). The shaded region represents 90% confidence interval.

It can be observed that the initial policy with CL is near-optimal, while the training without CL requires about 1 million episodes to acquire the optimal policy. Therefore, the application of CL can significantly accelerate the training of DRL.

D. Power Dispatching in the Upper Layer

Next, focusing on the operation of the upper layer, the power dispatching issue in the ER network depicted in Fig. 6 (MGs are omitted) is considered. Here, the ER network is interconnected to the main gird via ER_1 and ER_3. The parameters for the upper layer are shown in Table IV.

The maximum value of power exchange between MGs and ERs is set to be 200kW. For illustrative purpose, the electricity

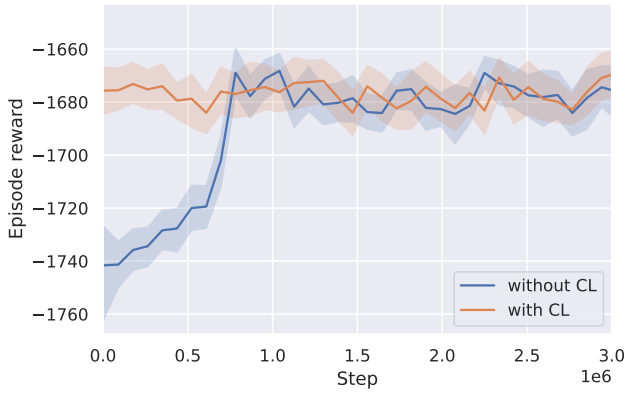


Fig. 5: The comparisons between DRL with CL and without CL. Note, DRL with CL converges much faster than DRL without CL while DRL with CL outperforms DRL without CL by a small margin.

TABLE IV: Parameters for upper layers.

Parameter	Value	Parameter	Value
μ_{12}	5×10^{-4}	$P^{ER}_{max,23}$	300
μ_{16}	5×10^{-4}	$P^{ER}_{max,34}$	300
μ_{23}	5×10^{-4}	$P^{ER}_{max,35}$	300
μ_{34}	5×10^{-4}	$P^{ER}_{max,45}$	300
μ_{35}	5×10^{-4}	$P^{ER}_{max,56}$	300
μ_{45}	5×10^{-4}	$\alpha^{p}_{1,min}$	0.4
μ_{56}	5×10^{-4}	$\alpha^{s}_{1,min}$	0.22
$P^{ER}_{max,12}$	300	$\alpha^{p}_{3,min}$	0.28
$P^{ER}_{max,16}$	300	$\alpha^{s}_{3,min}$	0.15

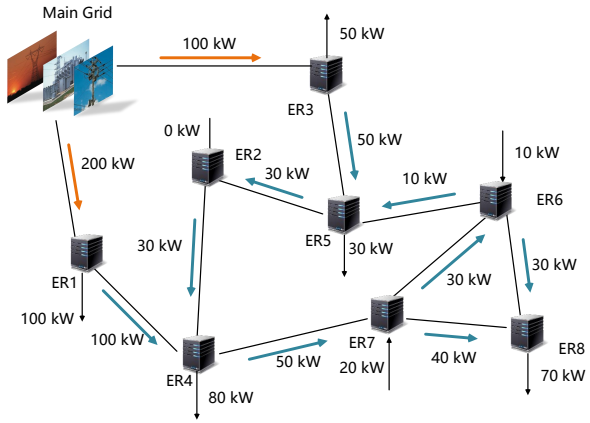


Fig. 6: The intuitive illustration of power dispatching in the upper layer. The orange arrows denote the power delivery between ERs and the main grid, the blue arrows denote the power dispatching between ERs, and the black arrows denotes the power exchange between ERs and MGs.

price is generated based on geometric Brownian motions, which is widely used in the research on stock prices. Based on the operation of the MGs in the bottom layer, the power exchange between ERs and MGs is determined. Then, the power dispatching scheme for the ER network is calculated via Algorithm 3, by which the energy transmission among ERs and the energy trading with the utility grid are determined.

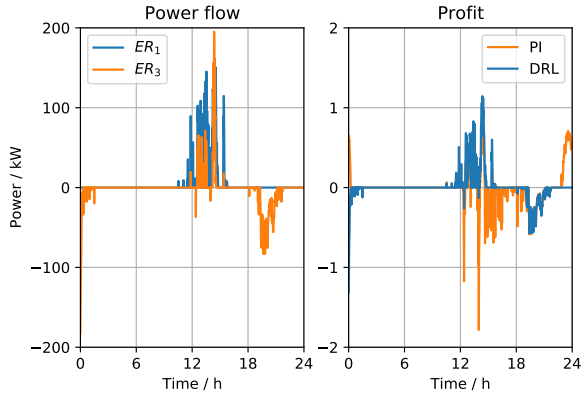


Fig. 7: (a) Power flow with the main grid; (b) The profit of ER network under DRL and PI methods.

According to the energy trading curves shown in Fig. 7, during $[0, 24h]$, the total amount of profit from energy trading under DRL method is 16.5 (currency unit is omitted), while that under PI method is 1.5. This shows the DRL method not only reduces the operation costs in the bottom layer, but also creates more profit via power trading in the upper layer.

Fig. 7 illuminates how the dynamical prices influence the power flows among ERs and the main grid in the upper layer. For the reason that the sale price at ER_1 is larger than that at ER_3, most surplus electricity is sold to the main grid through ER_1. Similarly, almost all electricity is purchased from the main grid through ER_3 due to its lower price.

VI. Conclusion

A layered bottom-up energy management scheme for EI has been proposed in this paper. DRL techniques have been applied to solve such a data-driven stochastic optimization problem. The simulation has shown that, under the two-layer architecture, compared with the conventional PI method and OPF method, our proposed method reduces generation cost by approximately 7.1% and 37%, and promotes the rational and efficient utilization of BESs. Besides, the introduced CL-based training strategy can significantly speed up the convergence and reduce the training cost during the training of DRL. Moreover, our method increases the profit of energy trading between ERs and the main grid.

In this paper, the weight coefficients of each component in cost functions (9) and (13) are assumed to be priorly known, which can be viewed as a limitation of this work. To achieve the bottom-up energy management mode, the formulated optimal control problems in both bottom layer and upper layer are indeed multi-objective optimization issues. For multiple sub-objectives to be optimized, how much their respective weights are appropriate is a key scientific problem to be studied in the future.

Appendix PSO-based PI control

When the conventional PSO method is applied, the PI controller of DGs is formulated as

$$\begin{aligned} u_i^g(t) =& -K_{i,1}^g P_i^{BES}(t) - K_{i,2}^g \left(\int_0^t P_i^{BES}(\tau)d\tau\right) \\ & -K_{i,3}^g (SOC_i(t) - SOC_i^{mid}) \\ & -K_{i,4}^g \left(\int_0^t (SOC_i(\tau) - SOC_i^{mid})d\tau\right), \end{aligned}$$

where $g \in \{\text{MT, FC, DEG}\}$, $K_{i,1}^g$, $K_{i,2}^g$, $K_{i,3}^g$ and $K_{i,4}^g$ are coefficients corresponding to P_i^{BES} and SOC_i for the proportional and integral terms, respectively. The PSO approach is applied to find the appropriate coefficients by minimizing the expectation of the cost function defined in (8). We define the position of one particle as

$$x = \{K_{i,j}^{MT}\}_{j=1}^4 \cup \{K_{i,j}^{FC}\}_{j=1}^4 \cup \{K_{i,j}^{DEG}\}_{j=1}^4$$

Note that the position x determines the PI control strategy, and also determines the expectation of J_i. Thus, we define the fitness value of position x as $f(x) = \mathbb{E}[J_i]$.

Algorithm 4 PSO for PI control of the i-th MG

for each particle $k = 1, \ldots, M$ **do**
  Randomly initialize position x_k and velocity v_k.
  Set personal best position $x_k^{pb} \leftarrow x_k$.
  Evaluate the fitness value $f(x_k)$ by Monte-Carlo method.
Set global postion $x^{gb} \leftarrow \text{argmin}_{x_k} f(x_k)$.
repeat
  for each particle $k = 1, \ldots, M$ **do**
    Update x_k and v_k:
    $v_k \leftarrow \omega^{PSO} \cdot v_k + c_1^{PSO} \cdot rand() \cdot \left(x_k^{pb} - x_k\right) + c_2^{PSO} \cdot rand() \cdot \left(x^{gb} - x_k\right)$
    $x_k \leftarrow x_k + v_k$
    Evaluate fitness value $f(x_k)$ by Monte-Carlo method.
    if $f(x_k) < f(x_k^{pb})$ **then**
      $x_k^{pb} \leftarrow x_k$
      if $f(x_k) < f(x^{gb})$ **then**
        $x^{gb} \leftarrow x_k$
until converge

In Algorithm 4, ω^{PSO}, c_1^{PSO} and c_2^{PSO} are the inertia weight factor, cognitive and social acceleration factors, respectively. The notation "argmin" stands for the argument of the minimum. Thus, $\text{argmin}_x f(x)$ refers to the point x that minimizes function $f(x)$. Besides, the random number of range (0,1) is denoted as $rand()$.

References

[1] A. Joseph and P. Balachandra, "Smart grid to energy Internet: A systematic review of transitioning electricity systems," IEEE Access, vol. 8, pp. 215787–215805, 2020.

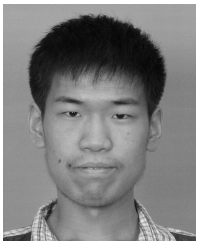
[2] H. M. Hussain, A. Narayanan, P. H. J. Nardelli, and Y. Yang, "What is energy Internet? Concepts, technologies, and future directions," IEEE Access, vol. 8, pp. 183127–183145, 2020.
[3] J. Cao, H. Hua, G. Ren, "Energy use and the internet," in: The SAGE Encyclopedia of the Internet, SAGE Publications, Thousand Oaks, CA, USA, 2018, pp. 344–350.
[4] K. Wang *et al.*, "A survey on energy Internet: architecture, approach, and emerging technologies," IEEE Syst. J., vol. 12, no. 3, pp. 2403–2416, 2017.
[5] C. Hao, Y. Qin, and H. Hua, "Energy "Routers", "Computers" and "Protocols"," in: A. Zobaa, J. Cao (eds), Energy Internet: Systems and Applications, Springer Nature Switzerland AG, 2020, pp. 193–208.
[6] H. Liang, H. Hua, Y. Qin, M. Ye, S. Zhang and J. Cao, "Stochastic Optimal Energy Storage Management for Energy Routers via Compressive Sensing," IEEE Trans. Ind. Inform., early access, DOI: 10.1109/TII.2021.3095141.
[7] M. Gao, K. Wang, and L. He, "Probabilistic model checking and scheduling implementation of an energy router system in energy Internet for green cities," IEEE Trans Ind. Inform., vol. 14, no. 4, pp. 1501–1510, 2018.
[8] R. Wang, J. Wu, Z. Qian, Z. Lin, and X. He, "A graph theory based energy routing algorithm in energy local area network," IEEE Trans. Ind. Inform., vol. 13, no. 6, pp. 3275–3285, 2017.
[9] K. Zhou, S. Yang, and Z. Shao, "Energy Internet: the business perspective," Appl. Energy, vol. 178, pp. 212–222, 2016.
[10] H. Hua, Y. Qin, Z. He, L. Li, and J. Cao, "Energy sharing and frequency regulation in energy Internet via mixed H_2/H_∞ control with Markovian jump," CSEE J. Power & Energy Syst., early access, DOI: 10.17775/CSEEJPES.2019.01900.
[11] L. K. Gan, P. Zhang, J. Lee, M. A. Osborne, and D. A. Howey, "Data-driven energy management system with Gaussian process forecasting and MPC for interconnected microgrids," IEEE Trans. Sust. Energy, vol. 12, no. 1, pp. 695–704, Jan. 2021.
[12] Y. Li, D. W. Gao, W. Gao, H. Zhang, and J. Zhou, "Double-mode energy management for multi-energy system via distributed dynamic event-triggered Newton-Raphson algorithm," IEEE Trans. Smart Grid, vol. 11, no. 6, pp. 5339–5356, Nov. 2020.
[13] Y. Li, D. W. Gao, W. Gao, H. Zhang, and J. Zhou, "A distributed double-Newton descent algorithm for cooperative energy management of multiple energy bodies in energy Internet," IEEE Trans. Ind. Inform., vol. 17, no. 9, pp. 5993–6003, Sept. 2021.
[14] F. Si, *et al.*, "Cost-efficient multi-energy management with flexible complementarity strategy for energy Internet," Appl. Energy, vol. 231, pp. 803–815, 2018.
[15] P. Kou, D. Liang, and L. Gao, "Distributed EMPC of multiple microgrids for coordinated stochastic energy management," Appl. Energy, vol. 185, pp. 939–952, 2017.
[16] R. Jing *et al.*, "Distributed or centralized? Designing district-level urban energy systems by a hierarchical approach considering demand uncertainties," Appl. Energy, vol. 252, pp. 113424, 2019.
[17] R. Mudumbai, S. Dasgupta, and B. B. Cho, "Distributed control for optimal economic dispatch of a network of heterogeneous power generators," IEEE Trans. Power Syst., vol. 27, no. 4, pp. 1750–1760, 2012.
[18] H. Hua, C. Hao, and Y. Qin, "Internet thinking for layered energy infrastructure," in: A. Zobaa, J. Cao (eds), Energy Internet: Systems and Applications, Springer Nature Switzerland AG, 2020, pp. 421–437.
[19] H. Pourbabak, J. Luo, T. Chen, and W. Su, "A novel consensus-based distributed algorithm for economic dispatch based on local estimation of power mismatch," IEEE Trans. Smart Grid, vol. 9, no. 27, pp. 5930–5942, 2018.
[20] R. K. Jain, J. Qin, and R. Rajagopal, "Data-driven planning of distributed energy resources amidst socio-technical complexities," Nature Energy, vol. 2, no. 8, pp. 17112, 2017.
[21] 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, 2019.
[22] B. W. Yi, J. H. Xu, and Y. Fan, "Inter-regional power grid planning up to 2030 in China considering renewable energy development and regional pollutant control: a multi-region bottom-up optimization model," Appl. Energy, vol. 184, pp. 641–658, 2016.
[23] P. N. Georgiou, "A bottom-up optimization model for the long-term energy planning of the Greek power supply sector integrating mainland and insular electric systems," Computers & Operations Research, vol. 66, pp. 292–312, 2016.
[24] S. Xia *et al.*, "A fully distributed hierarchical control framework for coordinated operation of DERs in active distribution power networks," IEEE Trans. Power Syst., vol. 34, no. 6, pp. 5184–5197, 2019.
[25] K. Lv *et al.*, "Hierarchical learning optimisation method for the coordination dispatch of the inter-regional power grid considering the quality of service index," IET Gen. Transm. & Distrib., vol. 14, no. 18, pp. 3673–3684, 2020.
[26] H. Hua, Y. Qin, C. Hao, and J. Cao, "Optimal energy management strategies for energy Internet via deep reinforcement learning approach," Appl. Energy, vol. 239, pp. 598–609, 2019.
[27] E. Mocanu *et al.*, "On-line building energy optimization using deep reinforcement learning," IEEE Trans. Smart Grid, vol. 10, no. 4, pp. 3698–3708, 2019.
[28] H. Zhang, J. Zhou, Q. Sun, J. M. Guerrero, and D. Ma, "Data-driven control for interlinked AC/DC microgrids via model-free adaptive control and dual-droop control," IEEE Trans. Smart Grid, vol. 8, no. 2, pp. 557–571, 2017.
[29] Y. Du and F. Li, "Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning," IEEE Trans. Smart Grid, vol. 11, no. 2, pp. 1066–1076, 2020.
[30] H. Xiao *et al.*, "A comparative study of deep neural network and meta-model techniques in behavior learning of microgrids," IEEE Access, vol. 8, pp. 30104–30118, 2020.
[31] Y. Ye, D. Qiu, X. Wu, G. Strbac, and J. Ward, "Model-free real-time autonomous control for a residential multi-energy system using deep reinforcement learning," IEEE Trans. Smart Grid, vol. 11, no. 4, pp. 3068–3082, 2020.
[32] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 26–38, 2017.
[33] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in Proc. 26th Annual Int. Conf. Machine Learning, Montreal, Quebec, Canada, Jun., 2009, pp. 41–48.
[34] Z. Qin, D. Liu, H. Hua, and J. Cao, "Privacy preserving load control of residential microgrid via deep reinforcement learning," IEEE Trans. Smart Grid, vol. 12, no. 5, pp. 4079–4089, Sept. 2021.
[35] Dataport, pecan street inc., https://dataport.cloud/.
[36] Q. Zang and L. Zhang, "Asymptotic behaviour of the trajectory fitting estimator for reflected Ornstein–Uhlenbeck processes," J. Theor. Prob., vol. 3, pp. 1–19, 2017.
[37] H. Hua, J. Cao, G. Yang, and G. Ren, "Voltage control for uncertain stochastic nonlinear system with application to energy Internet: Non-fragile robust H_∞ approach," J. Math. Anal. Appl., vol. 463, no. 1, pp. 93–110, 2018.
[38] B. Heymann *et al.*, "Continuous optimal control approaches to microgrid energy management," Energy Syst., vol. 9, no. 1, pp. 59–77, 2015.
[39] Z. Qin *et al.*, "Optimal electricity trading strategy for a household microgrid," in Proc. 16th IEEE Int. Conf. Control & Automation, Singapore, 2020, pp. 1308–1313.
[40] S. Borovkova and M. D. Schmeck, "Electricity price modeling with stochastic time change," Energy Economics, vol. 63, pp. 51-65, 2020.
[41] S. Narvekar and P. Stone, "Learning curriculum policies for reinforcement learning", in Proc.18th Int. Conf. Autonomous Agents & MultiAgent Systems, Richland, SC, pp. 25–33, 2019.
[42] S. Diamond and S. Boyd, "Cvxpy: A Python-embedded modeling language for convex optimization," J. Mach. Learn. Res., vol. 17, no. 1, pp. 2909–2913, 2016.
[43] L. Thurner *et al.*, "Pandapower—An open-source Python tool for convenient modeling, analysis, and optimization of electric power systems," IEEE Trans. Power Syst., vol. 33, no. 6, pp. 6510–6521, 2018.

Haochen Hua (M'16) was born in Jiangsu, 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.

He has published over 50 papers and has authored 3 books. His current research interests include energy Internet system modelling and optimization, optimal and robust control theory, and stochastic calculus.

Zhaoming Qin is currently pursuing his master degree at the Department of Automation, Tsinghua University, Beijing, China. He received the B.Sc. in Automation from Beihang University, Beijing, China, in 2019. His current research focuses on reinforcement learning, specifically in the context of smart grids.

Nanqing Dong is currently a PhD student at the Department of Computer Science, University of Oxford, Oxford, UK. He received his master degree from the Department of Statistical Science, Cornell University, Ithaca, NY, USA, in 2017. Prior to Oxford, he worked at the Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA. His research interests include machine learning, computer vision, optimization, and quantum computing.



Yuchao Qin was born in Henan, P. R. China in 1994. He received the B.Sc. degree in automation and the M.S. degree in control science and engineering from Tsinghua University, Beijing, P. R. China in 2017 and 2020, respectively.

He is currently a PhD student in the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK. His current research interests include control and optimization, reinforcement learning, inverse reinforcement learning, and their applications in healthcare.

Maojiao Ye received the B.Eng. degree in Automation from the University of Electronic Science and Technology of China, Sichuan, China, in 2012 and the Ph.D. degree from Nanyang Technological University, Singapore, in 2016. She is currently a Professor with the School of Automaton, Nanjing University of Science and Technology. Prior to her current position, she was a Research Fellow in the School of Electrical and Electronic Engineering at Nanyang Technological University from 2016-2017.

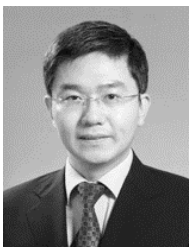
Dr. Ye was a recipient of Guan Zhao-Zhi Award in the 36th Chinese Control Conference 2017 (first author) and a recipient of the Best Paper Award in the 15th IEEE International Conference on Control and Automation 2019 (sole author). Her research interests include game theory, distributed optimization, and their applications.

Zidong Wang (SM'03-F'14) was born in Jiangsu, China, in 1966. He received the B.Sc. degree in mathematics in 1986 from Suzhou University, Suzhou, China, and the M.Sc. degree in applied mathematics in 1990 and the Ph.D. degree in electrical engineering in 1994, both from Nanjing University of Science and Technology, Nanjing, China.

He is currently Professor of Dynamical Systems and Computing in the Department of Computer Science, Brunel University London, U.K. From 1990 to 2002, he held teaching and research appointments in universities in China, Germany and the UK. Prof. Wang's research interests include dynamical systems, signal processing, bioinformatics, control theory and applications. He has published more than 600 papers in international journals. He is a holder of the Alexander von Humboldt Research Fellowship of Germany, the JSPS Research Fellowship of Japan, William Mong Visiting Research Fellowship of Hong Kong.

Prof. Wang serves (or has served) as the Editor-in-Chief for *International Journal of Systems Science*, the Editor-in-Chief for *Neurocomputing*, the Editor-in-Chief for *Systems Science & Control Engineering*, and an Associate Editor for 12 international journals including IEEE Transactions on Automatic Control, IEEE Transactions on Control Systems Technology, IEEE Transactions on Neural Networks, IEEE Transactions on Signal Processing, and IEEE Transactions on Systems, Man, and Cybernetics-Part C. He is a Member of the Academia Europaea, a Member of the European Academy of Sciences and Arts, an Academician of the International Academy for Systems and Cybernetic Sciences, a Fellow of the IEEE, a Fellow of the Royal Statistical Society and a member of program committee for many international conferences.

Xingying Chen was born in Wuxi, China. She received B.Eng. and Ph.D. degrees from Southeast University, China, in 1984 and 2002, respectively, and an M.Eng. degree from Hohai University, China, in 1995, all in electrical engineering. She joined the faculty of Hohai University, China, in 1984, and has become Professor there since 2002. Her current research interests include distribution and utilization system of electric power, power market, energy management and control, energy economics.



Junwei Cao (SM'05) received the bachelor's and master's degrees in control theories and engineering from Tsinghua University, Beijing, China, in 1998 and 1996, respectively, and the Ph.D. degree in computer science from the University of Warwick, Coventry, U.K., in 2001. He is currently a Professor of Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China.

He has authored/coauthored more than 200 papers and cited by international scholars for more than 18 000 times. He has authored or edited eight books. His research interests include distributed computing technologies and energy/power applications.

Dr. Cao is a Senior Member of the IEEE Computer Society and a member of the ACM and CCF.