Analysis of Evolutionary Dynamics for Bidding Strategy Driven by Multi-Agent Reinforcement Learning

Ziqing Zhu, *Student Member, IEEE*, Ka Wing Chan, *Member, IEEE*, Siqi Bu, *Senior Member, IEEE*,
Siu Wing Or, Xiang Gao, and Shiwei Xia, *Senior Member, IEEE*

***Abstract*—In this letter, the evolutionary game theory (EGT) with replication dynamic equations (RDEs) is adopted to explicitly determine the factors affecting energy providers' (EPs) willingness of using the market power to uplift the price in the bidding procedure, which could be simulated using the win-or-learn-fast policy hill climbing (WoLF-PHC) algorithm as a multi-agent reinforcement learning (MARL) method. Firstly, empirical and numerical connections between WoLF-PHC and RDEs is proved. Then, by formulating RDEs of the bidding procedure, three factors affecting the bidding strategy preference are revealed, including the load demand, severity of congestion, and the price cap. Finally, the impact of these factors on the converged bidding price is demonstrated in case studies, by simulating the bidding procedure driven by WoLF-PHC.**

***Index Terms*—market power, multi-agent reinforcement learning, evolutionary game theory, bidding strategy**

I. Introduction

THE increasing amount of bilateral energy trading has resulted in massive congestion of tie-lines and accounts for higher complexity of transmission pricing. Meanwhile, the deregulated electricity market paradigm endogenizes strategic interactions and bidding games among multiple stakeholders intending for chasing more profits [1]. Most importantly, some stakeholders will intentionally uplift their bidding price to manipulate the locational marginal price (LMP) for more remuneration. This is known as the "market power" [2], which adversely contribute to the maximization of social welfare. Existing literatures in the field of market force mitigation are all based on static bidding game model, while little research provides insights based on the simulation of dynamic bidding procedure among multiple EPs. Thus, they failed to reveal the inherent reason why the EPs are motivated to uplift their bidding price, even willing to take the risk of failure of bidding. By answering this question, the market operator would be enlightened on how to impose proper regulations using the incentive compatibility method [3] to mitigate such abuse of market power.

In [4], the WoLF-PHC algorithm, was adopted to simulate the dynamic bidding among multiple EPs and to compute the Nash equilibrium point. However, the factors affecting the converged bidding price are implicitly indicated therein. In [5], the numerical connection between EGT with RDEs and some baseline MARL algorithms is proved, implying that EGT with RDEs can explicitly reveal the factors affecting the converged result in MARL. Thus, in this letter, the correlation between WoLF-PHC and EGT is investigated and adopted to analyse the learning dynamics. Specifically, the contributions of this letter are outlined as follows:

1) The connection between WoLF-PHC based MARL and EGT with RDEs is firstly investigated.
2) RDEs of bidding among EPs considering congestion management are formulated to reveal indicators affecting EPs' converged bidding price driven by WoLF-PHC.
3) The Evolutionary Strategy Stability (ESS) analysis is conducted to analyse how these indicators will affect the learning dynamics, and conclusions are validated by case studies in a 2-bus test network and IEEE 14-Bus system.

II. Market Bidding and Clearing Model Considering Congestion Management

The interactive bidding and market clearing procedure is commonly formulated as a bi-level dynamic programming model. In this model, all EPs are considered as conventional generators with controllable power output. To capture the methodology of MARL and integrate the impact of congestion management into EPs' decision making, the conventional bi-level model is modified as follows with all well-known constraints neglected:

EPs:

$$Max\ \ R^t_{en,i}\big(P^t_{en,i},\lambda^t_{en,i} \mid \lambda^{t-1}_{en,mar,i},P^{t-1}_{en,allo,i}\big) - C^t_{en,i}(P^t_{en,i}) + R^t_{surp,i} \tag{1}$$

$$R^t_{surp,i} = \sum\nolimits_{j\in J}(\lambda^t_{en,i}-\lambda^t_{en,j})\mu_{ij}F^t_{ij} \tag{2}$$

Market Operator:

$$Min\ \ C^t_{MO,en}\big(\lambda^t_{en,mar,i},P^t_{en,allo,i} \mid P^t_{en,i},\lambda^t_{en,i}\big) \tag{3}$$

$$C^t_{MO,en} = \sum\nolimits_{i\in I}\lambda^t_{en,mar,i}P^t_{en,allo,i} \tag{4}$$

This work was jointly supported by the Research Grants Council of the HKSAR Government (Grant No. R5020-18) and the Innovation and Technology Commission of the HKSAR Government to the Hong Kong Branch of National Rail Transit Electrification and Automation Engineering Technology Research Center (Grant No. K-BBY1), and the National Natural Science Foundation of China (52077075).

Ziqing Zhu, and Siu Wing Or are with the Hong Kong Branch of National Rail Transit Electrification and Automation Engineering Technology Research Center, and Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China (e-mail: ziqing.zhu@connect.polyu.hk, derek.s.w.or@polyu.edu.hk).

Ka Wing Chan, Siqi Bu, and Xiang Gao are with the Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China (e-mail: eekwchan@polyu.edu.hk, siqi.bu@polyu.edu.hk, jocelyn. gao@connect.polyu. hk).

Shiwei Xia is with the State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources, School of Electrical and Electronic Engineering, the North China Electric Power University, Beijing, China (e-mail: s.w.xia@ncepu.edu.cn).

The objective function of the i^{th} EP at time slot t is to maximize the total revenue $R_{en,i}^{t}$, and decision variables are the available capacity $P_{en,i}^{t}$ and the bidding price $\lambda_{en,i}^{t}$, based on market clearing results $[\lambda_{en,mar,i}^{t-1}, P_{en,allo,i}^{t-1}]$, in which $\lambda_{en,mar,i}^{t-1}$ denotes the LMP. The congestion surplus $R_{surp,i}^{t}$ is the multiplication of LMP difference between connected area *i* and *j*, and the tie-line capacity is F_{ij}^{t}, while $\mu_{ij}F_{ij}^{t}$ denotes the pre-purchased amount of Financial Transmission Right (FTR) [2] capacity by the i^{th} provider. The objective function of the market operator is to minimize the total cost $C_{MO,en}^{t}$ by optimally allocate $P_{en,allo,i}^{t}$ to each EP.

III. Evolutionary Dynamics of EPs' Bidding Strategy

A. Connections between MARL and EGT

As elaborated in previous work [4], such dynamic bi-level model (1)-(4) can be solved by MARL algorithms, which are intended to optimize the so-called "policy" of each "agent". The policy is the function mapping the "state" and the "action", i.e. to compute the optimal action with maximum "reward" in each state. For EPs, the state refers to $[\lambda_{en,mar,i}^{t-1}, P_{en,allo,i}^{t-1}]$, the action is $[P_{en,i}^{t}, \lambda_{en,i}^{t}]$. The EGT with RDEs, instead, presents the change of probability of multiple "players" selecting different "strategies", and these players will imitate the strategy of those who obtain the largest "payoff" [5]. Empirically, the strategy can be considered as the principle of selecting actions.

The numerical connection between the EGT and MARL lies in the speed of change of strategies and actions. If the speed of "strategy change" can be proved to be proportional to that of "policy change", then the expression of policy change can be directly replaced by that of strategy change [5], in which the factors affecting the converged result are explicitly indicated.

From the perspective of EGT, the game is assumed to be a two-player, two-strategy model, in which the two players refer to EPs located in the congested lines with different bidding strategies, including price-taker (strategy 1) and price-maker [4] (strategy 2). For EPs taking the price-taker strategy, they tend to propose a relatively low bidding price, to secure the success of bidding. For EPs taking the price-maker strategy, they tend to submit a higher bidding price, to pursue more benefit while taking the risk of failure of bidding. Based on the derivations presented in [5], the change of probability of player x and y [10] selecting different strategies p and q can be written as:

$$\frac{dx_p^t}{dt} = x_p^t x_q^t [y_q^t \boldsymbol{\mathcal{W}} \boldsymbol{\mathcal{R}} \boldsymbol{\mathcal{W}}^T + \mathcal{R}_{p,q}^t - \mathcal{R}_{q,q}^t] \tag{5}$$

where x_p^t denotes probability of player x selecting strategy p, $x_q^t = 1 - x_p^t$, $y_q^t = 1 - y_p^t$, $\boldsymbol{\mathcal{W}} = (1, -1)$, and $\boldsymbol{\mathcal{R}}$ denotes the payoff matrix:

$$\boldsymbol{\mathcal{R}} = \begin{pmatrix} \mathcal{R}_{p,p}^t & \mathcal{R}_{p,q}^t \\ \mathcal{R}_{q,p}^t & \mathcal{R}_{q,q}^t \end{pmatrix} \tag{6}$$

In conventional WoLF-IGA algorithm [6], the updating speed of policy is the gradient of expected reward to the policy, based on the assumption that the underlying game and the converged policy is known, as formulated below:

$$\frac{\partial \mathbb{E}(x_p^t, x_q^t)}{\partial x_p^t} = \frac{\partial}{\partial x_p^t} \left\{ \Delta_{sa}(x_p^t, x_q^t) \boldsymbol{\mathcal{R}} \begin{pmatrix} y_p^t \\ y_q^t \end{pmatrix} \right\} \tag{7}$$

where Δ_{sa} denotes the learning coefficient [4]. The WoLF-PHC algorithm removes such assumption by approximating the equilibrium using the average policy updated in each iteration. Hence, their dynamics are essentially the same [6], which can be further simplified as follows:

$$\frac{\partial \mathbb{E}(x_p^t, x_q^t)}{\partial x_p^t} = \Delta_{sa} [y_p^t \boldsymbol{\mathcal{W}} \boldsymbol{\mathcal{R}} \boldsymbol{\mathcal{W}}^T + \mathcal{R}_{p,q}^t - \mathcal{R}_{q,q}^t] \tag{8}$$

It can be concluded from (5) and (8) that the speed of both the strategy change in RDEs and the policy change in WoLF-PHC is proportional to $[y_p^t \boldsymbol{\mathcal{W}} \boldsymbol{\mathcal{R}} \boldsymbol{\mathcal{W}}^T + \mathcal{R}_{p,q}^t - \mathcal{R}_{q,q}^t]$. Hence, it is reasonable to use RDEs for analyzing the factors affecting the bidding strategy of EPs.

B. RDEs of EPs

Derived from (5), the RDEs are then formulated as follows:

$$x_p^t = x_p^{t0} + \int \frac{dx_p^t}{dt}\, dt \tag{9}$$

$$\mathcal{R}_{1,1}^t = \lambda_{en,1}^t P_{en,allo,1}^t + (\lambda_{en,2}^t - \lambda_{en,1}^t)\mu_{ab} F_{ab}^t \tag{10}$$

$$\mathcal{R}_{1,2}^t = \lambda_{en,2}^t P_{en,allo,2}^t = \lambda_{en,2}^t \left(P_{en,req} - P_{en,allo,1}^t\right) \tag{11}$$

$$\frac{dx_1}{dt} = x_1^t(1 - x_1^t)(\mathcal{R}_{1,1}^t - \mathcal{R}_{1,2}^t) \tag{12}$$

$$\frac{dx_2}{dt} = x_2^t(1 - x_2^t)(\mathcal{R}_{1,2}^t - \mathcal{R}_{1,1}^t) \tag{13}$$

(9) indicates that the proportion of the EP selecting the p^{th} strategy at time t is based on the initial value at time t_0, as well as the proportion change from t_0 to t. Here, EPs with price-maker strategy are assumed to always submit higher price than those with price-taker strategy. Thus, the revenue of EPs selecting the price-taker strategy includes the energy trading payoff and the congestion surplus, while the revenue of EPs selecting the price-maker strategy constitutes the trading payoff only, as indicated in (10) and (11). In (12) and (13), the rate of change of EPs' proportion is derived, which will be used to conduct the ESS analysis.

IV. Evolutionary Strategy Stability (ESS) Analysis

RDEs describe the procedure of players searching for the optimal strategy with maximum revenue, and this procedure terminates when all the players are not motivated to change their strategies, which is referred as the state of ESS. Here, the following criteria [7] is introduced for subsequent discussions:

Theorem 1. The strategy x_p will reach to ESS if and only if $\frac{dx_p}{dt} = 0$ and $\frac{d^2x_p}{dt^2} < 0$.

According to the formulated RDEs (16) and (17), if $\frac{dx_1}{dt} = 0$, then $x_1 = 0, x_2 = 1$, and $\frac{d^2x_1}{dt^2} = \mathcal{R}_{1,1}^t - \mathcal{R}_{1,2}^t$. Thus, the price-taker strategy will reach to ESS if and only if:

1) $x_1 = 0, \mathcal{R}_{1,1}^t - \mathcal{R}_{1,2}^t < 0$

→ if $\lambda_{en,2}^t > \frac{\lambda_{en,1}^t(P_{en,allo,1}^t - \mu_{ab}F_{ab}^t)}{P_{en,req} - P_{en,allo,1}^t - \mu_{ab}F_{ab}^t}$, then the proportion of EPs selecting price-taker strategy will be zero, i.e. all the EPs will tend to select the price-maker strategy, thus they will constantly uplift their bidding price and the final converged price will be the highest limitation.

2) $x_1 = 1, \mathcal{R}_{1,1}^t - \mathcal{R}_{1,2}^t > 0$

→ if $\lambda_{en,2}^t < \frac{\lambda_{en,1}^t(P_{en,allo,1}^t - \mu_{ab}F_{ab}^t)}{P_{en,req} - P_{en,allo,1}^t - \mu_{ab}F_{ab}^t}$, then the proportion of EPs selecting price-taker strategy will be one, i.e. all the EPs will tend to select the price-taker strategy, thus they will constantly decrease their bidding price and the final converged price will be the lowest limitation.

It can be concluded from the above that, the bidding strategy preference is correlated with the severity of congestion, the imposed limitation of bidding price, and the total energy demand. As EPs with price-taker strategy will be allocated with more offers than normal in the condition of excessive load demand and tie-line congestion, they will be motivated to uplift the bidding price to pursue for more benefits. This will subsequently result in the converged bidding price reaches to the maximum allowance. From the perspective of market operator, a lower price cap is effective to mitigate the potential abuse of such market force.

V. Case Study

A. Test Scenario 1: 2-Bus Network

A 2-bus test system shown in Fig.1 is designed to validate the aforementioned conclusions. The bidding procedure of the two generators is simulated using the WoLF-PHC algorithm [4]. It shall be noted that in each time slot of actual power market, the EPs only need to submit the bid once. In the simulation, it is assumed that each iteration refers to one round of bidding in the same condition (i.e. the bidding price limitation, the tie-line capacity, the total energy demand, etc.). Hence, the simulation refers to how the EPs "learn" to submit an optimal bidding price and finally reach to the convergence in the same conditions.

In case 1, the impact of congested capacity on the converged bidding price is first demonstrated by imposing different line capacity. In Scenario A, the bidding strategy converges to the price-taker strategy. In Scenario B, due to the deducted 40MW of line capacity, the price-maker with expensive bidding price will be allocated with more capacity than that in Scenario A. Thus, they will be motivated to submit higher prices, and the expected revenue of being a price-maker will be more than that of price-taker. In case 2, the impact of load demand on the converged bidding price is presented. As the load increases, both G1 and G2 will tend to uplift the bidding price to the maximum limitation. In case 3, with the price cap decreased from 280$/MWh to 200$/MWh, generators will tend to lower their bidding price, because the expected revenue of being a price-taker will be more than that of price-maker.

The evolutionary learning dynamics of the above scenarios can be simulated using the Vensim software [8] by inputting the corresponding parameters to the formulated RDEs (9)-(13). The simulation results, i.e. the probability change of EPs selecting different biding strategies, are shown in Fig.3 (a)-(f), which correspond to the bidding procedure shown in Fig.2 (a)-(f). This further validates the relationship between EGT and MARL. It shall be noted that the "time" in the x-axis does not corresponds to the "time" in real world; instead, it represents a generalized time to measure the speed of probability change.

B. Test Scenario 2: IEEE 14-Bus Network

The applicability of aforementioned conclusions is further demonstrated using the IEEE 14-bus network with 5 EPs as a more complicated scenario. The network configuration, generator characteristics, the base load and peak load data are extracted from [9]. The parameter settings of different cases are summarized in Table II.

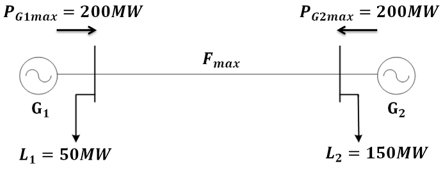

Fig.1. 2-Bus Test Network

TABLE I
Test Scenario 1: Parameters of Designated Cases

Case	Load Demand (MW)	Tie-line Capacity (MW)	Bidding Price Limitation ($/MWh)
1	100	Scenario A: 100 Scenario B: 80	[100,200]
2	Scenario C: 100 Scenario D: 200	100	[100,200]
3	100	100	Scenario E: [100,200] Scenario F: [100,280]

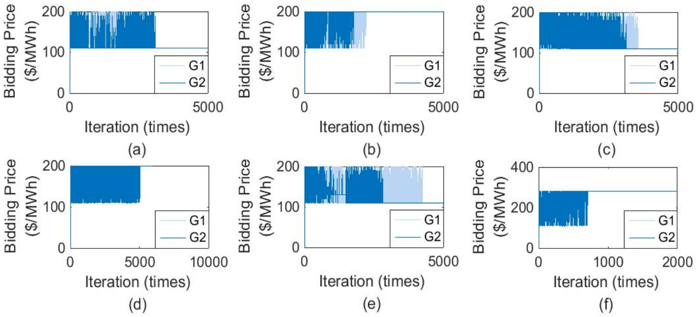

Fig.2 Biding Price Convergence in Different Scenarios

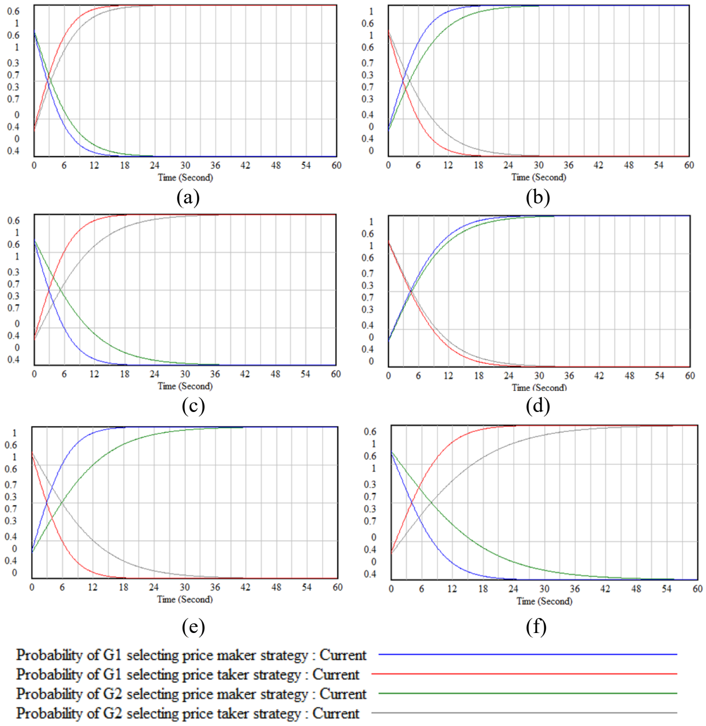

Fig.3. Evolutionary Dynamics in Different Scenarios

TABLE II
TEST SCENARIO 2: PARAMETERS OF DESIGNATED CASES

Case	Load Demand	Tie-line Capacity (MW)	Bidding Price Limitation ($/MWh)
4	Base Load	Rated	[100,200]
5	Base Load	Rated	[100,300]
6	Base Load	80% of Rating	[100,200]
7	Peak Load	Rated	[100,200]

The simulation results of bidding procedure are shown in Fig.4. In case 4, all EPs prefer the price-taker strategy under a lower demand and price cap. In case 5, due to the higher price cap, all EPs except for EP4 converge to the price-maker strategy, while EP4 tends to be the price-taker because of a cheaper generation cost. In case 6 and 7, compared with case 4, both EP3, EP4 and EP5 would uplift their bidding prices for more benefits due to the adjacent line congestion and demand increase, while EP1 and EP2 only slightly uplift the bidding price because the congestion in this area is not severe.

The applicability of the MARL (WoLF-PHC) method and EGT with RDEs can therefore be summarized as follows. The MARL is capable of simulating the bidding procedure among many EPs, but the key factors that will affect the final converged results remain unknown, which however can be addressed by formulating the RDEs and conducting the ESS analysis. Though it would be difficult to formulate the RDEs in the condition of many EPs, a simple 2-EP model is sufficient to identify the key factors. Therefore, the combination of MARL and EGT would be an effective and promising tool to simulate and analyze behaviors of participants in emerging market paradigms.

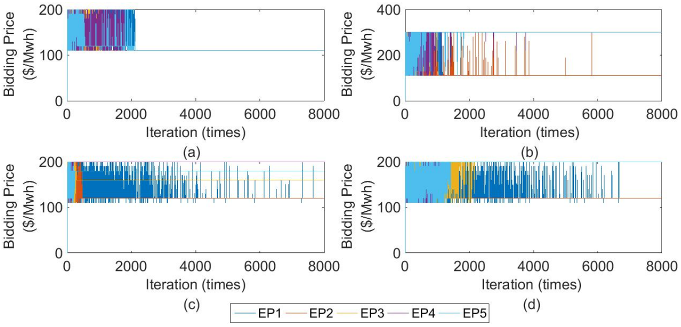

Fig.4. Biding Price Convergence in Different Cases

VI. CONCLUSION

In this letter, the EGT with RDEs is adopted to analyze the learning dynamics of EP's bidding strategy driven by MARL, based on the clarified empirical and numerical relationship between EGT and MARL. Three key factors that affect the converged bidding strategy are identified and analyzed. This methodology could be further extended to investigate behaviors of stakeholders in other emerging market paradigms.

REFERENCES

[1] Y. Ye, D. Qiu, M. Sun, D. Papadaskalopoulos, and G. Strbac, "Deep Reinforcement Learning for Strategic Bidding in Electricity Markets," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1343-1355, 2020.

[2] J. Nicolaisen, V. Petrov, and L. Tesfatsion, "Market power and efficiency in a computational electricity market with discriminatory double-auction pricing," *IEEE Trans. Evol. Comput*., vol. 5, no. 5, pp. 504-523, 2001.

[3] H. Guo and N. C. Yannelis, "Incentive compatibility under ambiguity," *Econ. Theory*, 2020.

[4] X. Gao, K. W. Chan, S. Xia, X. S. Zhang, K. Zhang, and J. Zhou, "A Multi-agent Competitive Bidding Strategy in a Pool-Based Electricity Market with Price-Maker Participants of WPPs and EV Aggregators," *IEEE Trans. Ind. Informatics*, vol. 3203, no. c, pp. 1-1, 2021.

[5] M. Keisers, Learning against Learning: Evolutionary Dynamics of Reinforcement Learning Algorithms in Strategic Interactions. 2012.

[6] A. Sherief and L. Victor, "A multiagent reinforcement learning algorithm with non-linear dynamics," *The Journal of artificial intelligence research*, vol. 33, pp. 521-549, 2008, doi: 10.1613/jair.2628.

[7] C. Alós-Ferrer and A. B. Ania, "The evolutionary stability of perfectly competitive behavior," *Econ. Theory*, vol. 26, no. 3, pp. 497-516, 2005.

[8] B. Arshad, N. Arshad, F. Mohamed, and N. Mohd, System dynamics : modelling and simulation. Singapore: Springer Singapore : Imprint : Springer, 2017.

[9] X. Fang, F. Li, Y. Wei, R. Azim, and Y. Xu, "Reactive power planning under high penetration of wind energy using Benders decomposition," *IET Gener. Transm. Distrib.*, vol. 9, no. 14, pp. 1835-1844, 2015.

[10] Z. Zhu, K. W. Chan, S. Bu, B. Zhou, and S. Xia, "Real-Time interaction of active distribution network and virtual microgrids: Market paradigm and data-driven stakeholder behavior analysis," *Appl. Energy*, vol. 297, p. 117107, 2021.

Ziqing Zhu (S'19) received the M.S. degree in electrical power systems engineering from the University of Manchester, UK, in 2019. He is now pursuing his Ph.D. with the Hong Kong Polytechnic University, Kowloon, Hong Kong SAR. His research interests include applications of game theory and AI techniques on energy markets and power system optimization.

Ka Wing Chan (M'98) received the B.Sc. (Hons) and Ph.D. degrees in electronic and electrical engineering from the University of Bath, U.K., in 1988 and 1992, respectively. He currently is an Associate Professor and Associate Head in the Department of Electrical Engineering of the Hong Kong Polytechnic University. His general research interests include power system stability, analysis and control, power grid integration, security, resilience and optimization, demand response management, etc.

Siqi Bu (S'11–M'12–SM'17) received the Ph.D. degree from the electric power and energy research cluster, The Queen's University of Belfast, Belfast, U.K., in 2012. Then he was with National Grid UK as an experienced UK National Transmission System Planner and Operator. He is an Associate Professor with The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, and also a Chartered Engineer with UK Engineering Council, London, U.K. His research interests include power system stability analysis and operation control, including wind power generation, PEV, HVDC, FACTS, ESS, and VSG.

Siu Wing Or received the B.Sc., M.Phil., and Ph.D. degrees in engineering physics from The Hong Kong Polytechnic University, Hong Kong, in 1995, 1997, and 2001, respectively. He is currently a Full Professor with the Department of Electrical Engineering, the Director of the Smart Materials and Systems Laboratory, and the Director of the Electrical Protection and High Voltage Coordination Laboratory, at The Hong Kong Polytechnic University. His research interests include smart materials and devices in the bulk, microscale, and nanoscale, condition and structural health monitoring, energy harvesting and storage, electromagnetics, ultrasonics, and vibration control

Xiang Gao received the B.Eng. degree in electrical engineering from Taiyuan University of Science and Technology, Taiyuan, China, in 2010, the M.Sc. degrees from Tianjin University of Technology, Tianjin, China and Newcastle University, Newcastle upon Tyne, UK, and the Ph.D. degree at the Hong Kong Polytechnic University, Hong Kong SAR, both in electrical engineering in 2014, 2015 and 2021 respectively. Her research interests include renewable energy resources, stochastic optimization and electricity market.

Shiwei Xia (S'11-M'15-SM'20) received the Ph.D. degree in power systems from The Hong Kong Polytechnic University, Hong Kong SAR, in 2015. Currently, he is with the State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources, School of Electrical and Electronic Engineering, North China Electric Power University, Beijing, China. His general research interests include security and risk analysis for power systems with renewable energies, distributed optimization and control of multiple sustainable energy sources in smart grid.