Received August 13, 2019, accepted August 28, 2019, date of publication September 6, 2019,
date of current version September 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2940005

Multi-Period and Multi-Spatial Equilibrium Analysis in Imperfect Electricity Markets: A Novel Multi-Agent Deep Reinforcement Learning Approach

YUJIAN YE[1,2], (Member, IEEE), DAWEI QIU[2], (Student Member, IEEE), JING LI[2], (Student Member, IEEE), AND GORAN STRBAC[2], (Member, IEEE)

[1]Fetch.AI, Milton, Cambridge CB4 0WS, U.K.
[2]Department of Electrical and Electronic Engineering, Faculty of Engineering, Imperial College London, London SW7 2AZ, U.K.

Corresponding author: Dawei Qiu (d.qiu15@lmperial.ac.uk)

ABSTRACT Previously works on analysing imperfect electricity markets have employed conventional game-theoretic approaches. However, such approaches necessitate that each strategic market player has full knowledge of the operating parameters and the strategies of its rivals as well as the computational algorithm of the market clearing process. This unrealistic assumption, along with the modeling and computational complexities, renders such approaches less applicable for conducting practical multi-period and multi-spatial equilibrium analysis. This paper proposes a novel multi-agent deep reinforcement learning (MA-DRL) based methodology, combining multi-agent intelligence, the deep policy gradient (DPG) method, and an innovative long short term memory (LSTM) based representation network for optimizing the offering strategies of multiple self-interested generation companies (GENCOs) as well as exploring the market outcome stemming from their interactions. The proposed approach is tailored to align with the nature of the examined problem by posing it, for the first time, in multi-dimensional continuous state and action spaces, enabling GENCOs to receive accurate feedback regarding the impact of their offering strategies on the market clearing outcome, and devise more profitable bidding decisions by exploiting the entire action domain, and thereby facilitates more accurate equilibrium analysis. The proposed LSTM-based representation network extracts discriminative features which further improves the learning performance and thus promises more profitable offerings strategies for each GENCO. Case studies demonstrate that the proposed method i) achieves a significantly higher profit than state-of-the-art RL methods for a single GENCO's optimal offering strategy problem and ii) outperforms the state-of-the-art equilibrium programming models in efficiently identifying an imperfect market equilibrium with / without network congestion. Quantitative economic analysis is carried out on the obtained equilibrium.

INDEX TERMS Deep neural networks, deep reinforcement learning, electricity markets, equilibrium programming, imperfect competition, multi-agent intelligence, strategic offering.

NOMENCLATURE

A. INDICES AND SETS

$t \in T$	Index and set of trading days.
$h \in H$	Index and set of hours.
$n, m \in M$	Indexes and set of nodes.
M_n	Set of nodes connected to node n through a transmission line.
$i \in I$	Index and set of generation companies (GENCO).
$i-$	Index of GENCOs other than i.
$j \in J$	Index and set of demands.
I_n, J_n	Set of GENCOs and demands connected to node n.
$b \in B$	Index and set of generation blocks.

B. PARAMETERS

N_H	Length of market horizon.
$\overline{F}_{n,m}$	Cpacity of transmission line (n, m) (MW).

The associate editor coordinating the review of this manuscript and approving it for publication was Qiuhua Huang.

$x_{n,m}$ Reactance of transmission line (n, m) (p.u.).
$\lambda^{G}_{i,b}$ Marginal cost of block b of GENCO i (£/MWh).
$\overline{o}_i$ Upper limit of strategic offering variable of GENCO i.
$\overline{g}_{i,b}$ Maximum power output limit of block b of GENCO i (MW).
R^U_i , R^D_i Ramp up / down limit of GENCO i (MW).
$D_{j,h}$ Power input of demand j at hour h (MW).

C. VARIABLES

$\theta_{n,h}$ Voltage angle at node n and period h (rad).
$o_{i,h}$ Strategic offering variable of GENCO i at hour h.
$g_{i,h,b}$ Power output of block b of GENCO i at hour h (MW).
$\lambda_{n,h}$ Locational marginal price at node n hour h (£/MWh).

I. INTRODUCTION

A. BACKGROUND AND MOTIVATION

The main motivation behind the deregulation of the electricity industry involves the unbundling of vertically integrated utilities and the introduction of competition in the generation and supply sectors of the industry in order to reduce the total system costs [1]. However, electricity markets are still characterized by a small number of large players. Therefore, these markets are better described as imperfectly rather than perfectly competitive. In this setting, market players do not necessarily act as *price-takers*. In particular, generation companies (GENCOs) occupying a large share of the market and / or strategically located in the transmission network are able to manipulate the electricity prices and increase their profits beyond the competitive equilibrium levels, through strategic offering. In other words, they act as *price-makers* and do not reveal their actual operating characteristics in their offers to the market but rather misreport them to increase their economic profits. This results in adverse consequences including increased prices and loss of market efficiency [1], [2].

Game-theoretic modeling approaches constitute the most common ones in the literature to study imperfect electricity markets. These approaches, depending on the modeling of one GENCO's behavior in relation to the behavior of his competitors, can be broadly classified into two categories: i) single GENCO's optimization models, which neglect the strategic interaction with the other GENCOs (i.e. treating the latter's strategies as fixed parameters) and ii) equilibrium programming models, which take into consideration the strategic interactions of all GENCOs.

In the first category, the decision-making process of a single strategic GENCO is usually modelled through a *bi-level optimization model* [3]–[5] which captures the interaction between the strategic player (modelled in the upper level (UL)) and the competitive clearing of the market (modelled in the lower level (LL)). Bi-level optimization problems are usually solved after transforming them to single-level *mathematical programs with equilibrium constraints* (MPEC), through the replacement of the LL problem by its equivalent *Karush-Kuhn-Tucker* (KKT) optimality conditions. An alternative approach to the above problem is to model the market clearing price at each hour as a function of the demand using a price-quota curve (or an inverse demand curve) [6]. However, the parameters of this function are determined based on exogenous data and therefore cannot accurately capture the impact of GENCOs' offering strategies on the formation of the market clearing prices, as opposite to the case with bi-level optimization models where the prices are endogenously determined in the LL problem.

In the second category, equilibrium programming models are employed when each GENCO takes into account the strategic behavior of its competitors. Such models aim at analysing the market outcome stemming from the interactions of multiple price-making GENCOs. The *Bertrand* (for modeling price game) [7], *Cournot* (for modeling quantity game) [8] and *supply function equilibrium* (SFE) models [9] constitute different imperfect equilibrium models reported in the literature. Furthermore, different computational techniques have been developed for computing the imperfect equilibrium. Authors in [10]–[13] formulate the problem by replacing each GENCO's MPEC problem by its KKT optimality conditions and and concatenate them together, resulting in a set of nonlinear constraints known as *equilibrium problem with equilibrium constraints* (EPEC). An *iterative diagonalization algorithm* (DIAG) is used in [14]–[17] to identify the imperfect equilibrium, in which each GENCO solves its own MPEC problem treating the strategies of the rest of the GENCOs as fixed, until the algorithm converges to a fixed market outcome. Furthermore, authors in [18] introduce the concept of *extremal market equilibrium* and formulate it into a mixed integer linear program (MILP) which provides an approximation of original EPEC problem.

Despite the theoretical soundness of the conventional game-theoretic modeling approaches, they suffer from several drawbacks. First of all, the inherent non-convexities and non-linearities presented in these models (due to the vast number of complementarity conditions and the mixed-integer linearization of some bilinear terms in these models [19]) render them very hard and computationally expensive to solve. Furthermore, such modeling and computational challenges are exacerbated in the multi-period and network-constrained framework investigated in this paper since the number and dimension of the decision variables (and therefore the complexity of the optimization problems) are increased considerably on the account of modeling these practical aspects of the market. Secondly, such approaches assume that GENCOs have full knowledge of the operating parameters and the strategies of its competitors as well as the computational algorithm of the market clearing process, which generally constitute a very limiting and unrealistic assumption. Lastly, such approaches discard the benefits (or the accumulated experiences) of learning from GENCOs' repeated (daily) interactions with the market clearing process [20].

Driven by the extensive complexity of the electricity markets and the high importance for a competitive economy, significant efforts have been made in developing new modeling approaches to facilitate more efficient and accurate equilibrium analysis. A very promising one of which is the *agent-based* and *reinforcement learning* (RL) approach, which recently has attracted increasing research attention, driven by the rapid advancements in artificial intelligence. In this modeling framework, strategic GENCOs (agents) are capable of learning their optimal strategies (actions) by utilizing experiences acquired from repeated interactions with the market clearing process (environment). In other words, GENCOs do not rely on any knowledge of the computational algorithm of the market clearing process and the operating parameters and offering strategies of their competitors, but only on their own operating parameters, the observed market clearing outcomes (e.g. the clearing prices and dispatches) and the publicly available information on the market condition (e.g. the load forecasts). Furthermore, such approaches avoid the significant modeling and computational complexity posed by traditional equilibrium programming models.

In this area, previous works [21]–[29] have employed conventional Q-learning algorithm and its variants [30]. This type of algorithms, however, suffers severely from the *curse of dimensionality* since it relies on look-up tables to approximate the action-value function for each possible state-action pair. This necessitates that the learning problem being set up in discrete state and action spaces, rendering it intractable as the number of possible states / actions grows large or their spaces are continuous. In the examined market problem, however, states of the environment and agents' actions are not only continuous, but also multi-dimensional (due to the multi-period nature of the problem). In this context, naïve discretization of the state space significantly reduces the accuracy of the state representation of the environment, distorting the feedback GENCOs receive regarding the impact of their offering strategies on the clearing outcome, On the other hand, naïve discretization of the action space may adversely change the feasible action domain, leading to sub-optimal offering strategies. Furthermore, this issue associated with single GENCO's optimization problem may also adversely affect the determination of the market equilibrium as the latter takes into account the interaction of multiple strategic GENCOs, rendering the respective equilibrium analysis less meaningful.

In the context of addressing such dimensionality challenges, authors in [31] proposed the *deep Q network* (DQN) method which employs a deep neural network (DNN) to approximate the action-value function, and has achieved expert human-level performance in playing Atari 2600 games. However, although previous work has validated good performance of the DQN method in problems with continuous state spaces, it exhibits less satisfactory performance in problems with continuous action spaces since the employed DNN is trained to produce discrete action-value estimates rather than continuous actions [32]. This significantly impedes its effectiveness in tackling the examined market problem, since the GENCOs' actions are continuous and multi-dimensional.

B. SCOPE AND CONTRIBUTIONS

This paper aims at addressing the limitations of state-of-the-art game-theoretic and RL methods by proposing a novel *multi-agent deep reinforcement learning* (MA-DRL) based methodology, namely, the *deep policy gradient* (DPG) method with an innovative *Long-short Term Memory* (LSTM) based representation network, for optimizing the offering strategies of multiple self-interested GENCOs as well as exploring the market outcome stemming from their interactions. Case studies demonstrate the value of the proposed methodology by comparing it against state-of-the-art game-theoretic and RL methods in facilitating *multi-period, multi-spatial market equilibrium analysis*.

More specifically, the novel contributions of this paper are outlined below:

- A novel MA-DRL based methodology, namely MA-DPG-LSTM method, combining multi-agent intelligence and a DPG-LSTM method, is developed to address the examined problem. The proposed approach is tailored to align with the nature of the examined problem by establishing it in multi-dimensional continuous state and action spaces, enabling strategic GENCOs to receive accurate feedback regarding the impact of their bidding decisions on the market clearing outcome, and devise more profitable bidding decisions by exploiting the entire action domain. To the best of the authors' knowledge, this is the first time that an equilibrium programming problem is addressed with the consideration of both multi-dimensional continuous state and action spaces using a MA-DRL based approach.

- An LSTM-based representation network is proposed to extract discriminative features from raw data on the market condition and clearing outcome, which contributes to enhancing the learning performance of the proposed method. Furthermore, an experience reply buffer has been proposed to break the temporal correlations existed in the consecutively generated training samples and enhance sampling efficiency.

- Case studies on a test market with day-ahead horizon and hourly resolution, operating over the IEEE Reliability Test System demonstrate that, for a single GENCO's optimal offering strategy model, the proposed method achieves a significantly higher profit than state-of-the-art RL methods (Q-learning, DQN, and DPG) and approximates very closely the profit obtained by the conventional bi-level/MPEC approach which provides the benchmark solution.

- For the computation of the imperfect market equilibrium, case studies demonstrate that the proposed method outperforms state-of-the-art equilibrium programming models (EPEC, diagonalization, and MILP approaches) in efficiently identifying a multi-period and / or multi-spatial imperfect market equilibrium. Quantitative economic analysis is conducted on the obtained equilibrium in the absence / presence of network congestion.

C. PAPER STRUCTURE

The rest of this paper is organized as follows. Section II presents the formulation of examined market modeling problem. Section III details the proposed MADRL-based methodology. Section IV presents case studies validating the proposed methodology. Section IV presents case studies validating the proposed methodology. Finally, Section V discusses conclusions and future work of this work.

II. MARKET MODELING PROBLEM FORMULATION

A. PROPOSED MULTI-AGENT MARKET ARCHITECTURE

The examined market is modeled as a multi-agent system with GENCOs as agents. Before a trading day t begins, the market operator (MO) announces the 24-hour load forecast for day $t+1$. On day t, GENCOs are required to submit their supply offers to the MO. Based on the collected supply offers, the MO performs the market clearing (refers to the market clearing model presented in Section II-C). Subsequently, the MO publishes the market clearing outcome comprising of locational marginal prices (LMP) and generation dispatches to the GENCOs.

B. GENERATION COMPANY MODEL

For clarity reasons and without loss of generality, we assume that each GENCO owns a single generation unit. However, the model can be readily extended to allow GENCOs owing multiple generation units. The variable production cost of GENCO i at hour h is represented by a piece-wise linear cost function as:

$$C_{i,h,b}(g_{i,h,b}) = \lambda^{G}_{i,b} g_{i,h,b} \tag{1}$$

By taking the derivative on both side of (1), the marginal cost (2) expresses the step-wise offer curve (consisting of a number of blocks) that GENCO i submits to the market at each trading day.

$$MC_{i,h,b}(g_{i,h,b}) = \lambda^{G}_{i,b} \tag{2}$$

GENCOs generally exercise market power through either submitting offers higher than their true marginal costs (i.e. *economic withholding*) or offering less that than their true generation capacity (i.e. *physical withholding*) to the market [2]. In this paper, GENCO can exercise market power considering a combination of both economic and physical withholding, in which case the strategic marginal cost function is expressed by (3), where the value of the decision variable $1 \le o_{i,h} \le \overline{o}_i, \forall h$ represents the strategic behavior of GENCO i at hour h.

$$MC^{s}_{i,h,b}(g_{i,h,b}) = o_{i,h}\lambda^{G}_{i,b} \tag{3}$$

If $o_{i,h} = 1$, GENCO i behaves non-strategically and reveals its true marginal costs $\lambda^{G}_{i,b}, \forall b$ to the MO at hour h. Otherwise, if $1 < o_{i,h} \le \overline{o}_i$, GENCO i behaves strategically and reveals higher than its true marginal costs $(o_{i,h} * \lambda^{G}_{i,b}, \forall h, \forall b)$ to the market at hour h. GENCO i should determine the value of $o_{i,h}$ at hour h by optimally trading off higher market prices and lower electricity production. In other words, a higher $o_{i,h}$ contributes to increasing market prices at h, but at the same time it contributes to decreasing the quantity sold by GENCO i at h, since GENCOs with lower submitted offers may replace i in the merit order. The DRL method presented in Section III-E provides an effective tool for individual price-maker GENCO to learn an optimal offering strategy $o_{i,h}$ from its repeated interactions with the market clearing process, based solely on its own operating parameters and the publicly available information announced by the MO.

In a multi-agent context, multiple self-interested profit-driven GENCOs tend to behave non-cooperatively with a target of exercising their individual market power, the nature of market competition in this context is *oligopoly* (as each GENCO usually owns a relatively large market share) and can be modeled as a *non-zero-sum stochastic game* [22], for which the underlying state transition is a Markov Chain and can be modeled as a *Markov Decision Process* (MDP).

C. MARKET OPERATOR MODEL

The modeled market is a pool-based, energy-only market with a day-ahead horizon and hourly resolution. Following the model employed in [4]–[6], [9]–[12], [14]–[18], [22], [24]–[26], [28], [29], the market is cleared through the solution of an network-constrained economic dispatch problem (4)-(11) target at minimizing the total generation cost; in order to account for the effect of the transmission network, the market clearing process incorporates a DC power flow model which yields LMP $\lambda_{n,h}$ for each node n and hour h.

$$\min_{V^{LL}} \sum_{i,h,b} o_{i,h}\lambda^{G}_{i,b} g_{i,h,b} \tag{4}$$

where

$$V^{LL} = \{g_{i,h,b}, \theta_{n,h}\} \tag{5}$$

subject to:

$$\sum_{j\in J_n} D_{j,h} - \sum_{i\in I_n,b} g_{i,h,b} + \sum_{m\in M_n} \frac{\theta_{n,h} - \theta_{m,t}}{x_{n,m}} = 0, \forall n, \forall h \tag{6}$$

$$0 \le g_{i,h,b} \le \overline{g}_{i,b}, \forall i, \forall h, \forall b \tag{7}$$

$$-R^{D}_{i} \le \sum_{b} g_{i,h,b} - \sum_{b} g_{i,(h-1),b} \le R^{U}_{i}, \forall h < N_H \tag{8}$$

$$-\overline{F}_{n,m} \le \frac{\theta_{n,h} - \theta_{m,h}}{x_{n,m}} \le \overline{F}_{n,m}, \forall n, \forall m \in M_n, \forall h \tag{9}$$

$$-\pi \le \theta_{n,h} \le \pi, \forall n, \forall h \tag{10}$$

$$\theta_{1,h} = 0, \forall h \tag{11}$$

The MO's objective (4) is to minimize the *perceived* total system production costs as revealed by the GENCOs' supply offers. This optimization is subject to nodal demand-supply balance constraints (6) (the dual variables of which constitute the LMPs at each node and each time period), generation capacity limits (7), time-coupling ramp rate constraints (8). Limits on transmission line capacities and voltage angles of nodes are enforced through constraints (9)-(10), respectively. Finally, constraint (11) identifies $n = 1$ as the reference node.

III. PROPOSED MULTI-AGENT DEEP REINFORCEMENT LEARNING METHODOLOGY

A. RL BACKGROUND

As discussed in Section II, the market outcome with competing strategic GENCO agents is oligopolistic and can be modeled as a MDP in which the learning behavior of each agent is governed by an RL algorithm. In this context, each adaptive agent interacts with a stochastic environment by sequentially selecting actions over a sequence of time steps, in order to maximize a cumulative reward.

Before introducing the proposed methodology, the preliminaries of MDP and RL are first presented in this section. An MDP compromises: 1) a state space $\mathcal{S}$; 2) an action space $\mathcal{A}$; 3) a transition dynamics distribution with conditional transition probability $p(s_{t+1}|s_t, a_t)$, satisfying the *Markov property*, i.e., $p(s_{t+1}|s_t, a_t) = p(s_{t+1}|s_1, a_1, ..., s_t, a_t)$ in state-action spaces; and 4) a reward function r: $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

The decision as to which action a_t is chosen in a certain state s_t is governed by a stochastic policy π: $\mathcal{S} \to \mathcal{P}(\mathcal{A})$, where $\mathcal{P}(\mathcal{A})$ is a set of probability measures on $\mathcal{A}$ and $\pi(a_t|s_t)$ is the conditional probability at a_t associated with the policy. The agent employs its policy to interact with the MDP and emit a trajectory of states, actions and rewards: $s_1, a_1, r_1, ..., s_T, a_T, r_T$ over $\mathcal{S} \times \mathcal{A} \times \mathbb{R}$. The return $R_t = \sum_{l=t}^{T} \gamma^{(l-t)} r(s_l, a_l)$ is the total discounted reward from time-step t onwards, where $\gamma \in [0, 1]$ is the discount factor that is used to trade off the importance between immediate and future rewards. The agents' goal through RL is to form an optimal policy that maximises the cumulative discounted reward from the start state $t = 1$, denoted by the *performance function* $\mathcal{J}(\pi) = \mathbb{E}[R_1|\pi]$, then we can write it as an expectation:

$$\begin{aligned}\mathcal{J}(\pi) &= \int_{\mathcal{S}} \rho^{\pi}(s) \int_{\mathcal{A}} \pi(a|s) r(s, a) \mathrm{d}a\mathrm{d}s \\ &= \mathbb{E}_{s\sim\rho^{\pi}, a\sim\pi}[r(s, a)] \end{aligned} \tag{12}$$

where $\rho^{\pi}(s)$ denotes the discounted state distribution governed by the policy π.

B. RL FORMULATION OF THE MARKET MODELING PROBLEM

In this section, we detail the RL formulation of the examined market modeling problem, the key elements associated with which are outlined as follows:

1) ***Agent***: Each strategic GENCO i constitutes the agent.

2) ***Environment***: The environment is represented by the day-ahead market clearing algorithm carried out the MO, as formulated in the optimization problem (4)-(11).

3) ***State***: The state vector $s_{i,t}$ serves as a feedback signal for GENCO i regarding the influence of its offering strategy on the status of the environment and is comprised of the market clearing outcome for trading day $t-1$ and the load forecast of day $t+1$ (both information is publicly available to GENCO i on day t). Specifically, $s_{i,t} = [g_{i,1:N_H}, \lambda_{i,1:N_H}, d_{i,1:N_H}] \in \mathcal{S}_i$ is a $3 \times N_H$-dimensional continuous vector where $g_{i,1:N_H} \in [0, \sum_b \overline{g}_{i,b}]$ and $\lambda_{(n:i\in I_n),1:N_H} \in [0, \lambda^{max}]$ represent, respectively, the generation dispatches of GENCO i and the LMPs for day $t-1$; and $d_{1:N_H} = \sum_j D_{j,1:N_H}$ denotes the total system demand forecast announced by MO for day $t+1$.

4) ***Action***: The action $a_{i,t} = [o_{i,1:N_H}] \in \mathcal{A}_i$ (encoded in output layer of the proposed DPG network (Fig. 1)) of GENCO i is a N_H-dimensional continuous vector where $o_{i,h} \in [1, \overline{o}_i]$ represents the N_H strategic offering decisions of GENCO i submitted to MO on day t.

5) ***Reward***: The reward $r_{i,t}$ of the GENCO i resultant from its offering strategy $a_{i,t}$ is set to be its economic profit pro_i (13), given by the difference between its revenue in the market and its operating cost.

$$pro_i = \sum_{h,b} (\lambda_{(n:i\in I_n),h} g_{i,h,b} - \lambda^{G}_{i,b} g_{i,h,b}) \tag{13}$$

C. BENCHMARK RL ALGORITHMS

1) Q-LEARNING

A popular method for RL is to make use of the *action value function* (or the *Q-value function*) $Q^{\pi}(s, a) = \mathbb{E}[R_1|s_1 = s, a_1 = a; \pi]$ which forms an estimation of the expected total discounted reward given an action a_t, at state s_t, and following

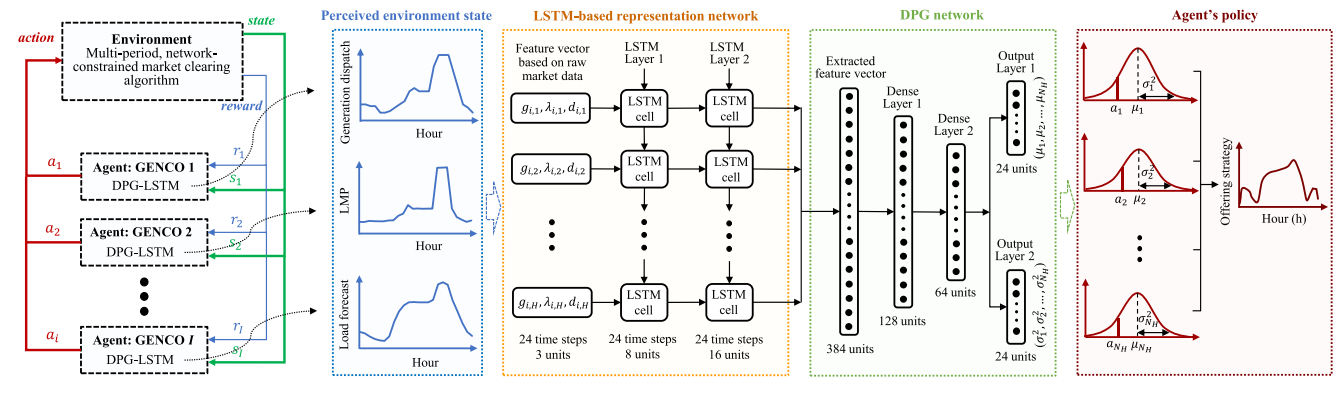


FIGURE 1. Workflow of proposed DPG-LSTM method.

the policy π from the succeeding states onwards. An optimal policy can be derived from the optimal Q-values $Q^*(s_t, a_t) = \max_\pi Q^\pi(s_t, a_t)$ by selecting the action corresponding to the highest Q-value in each state. The Q-value function can be described as a recursive format according to the *Bellman equation* [30]:

$$Q(s_t, a_t) = \mathbb{E}[r_t + \gamma Q(s_{t+1}, \pi(a_{t+1}|s_{t+1}))] \tag{14}$$

The Bellman equation indicates that the action value function under the current policy can be decomposed in terms of itself. Namely, Q-value can be updated by *bootstrapping*, i.e. we can improve the estimate of Q by using the current estimate of Q through *dynamic programming*. This serves as the foundation of Q-learning [33], a form of temporal difference (TD) learning [30]. The update of Q-value after taking action a_t in state s_t and observing the reward r_t and resulting state s_{t+1} is:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha\delta_t \tag{15}$$

$$\delta_t = r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \tag{16}$$

where $\alpha \in [0, 1]$ is the step size, δ_t represents the correction for the estimation of the Q-value function (known as the TD error), and $r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1})$ represents the target Q-value at time step t.

2) DEEP Q-NETWORK

To address the curse of dimensionality of Q-learning in multi-dimensional continuous state space (Section I-A), the DQN method [31] employs a DNN, parameterized by θ, which takes as input a continuous state s_t and outputs an estimate for the Q-value function (i.e. $Q(s_t, a_t) \approx Q(s_t, a_t|\theta)$) for each discrete action and, when acting, selects the maximally valued output at a given state. The training of the DNN is based on minimizing the following loss function representing the mean-squared TD error:

$$\mathcal{L}(\theta) = \mathbb{E}[\big(r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}|\theta) - Q(s_t, a_t|\theta)\big)^2] \tag{17}$$

D. PROPOSED DEEP POLICY GRADIENT NETWORK

Although DQN method has good performance in problems with continuous state spaces, its performance in problems with continuous action spaces is not satisfactory because the employed DNN is trained to produce discrete action-value estimates rather than continuous actions, which significantly hinders its effectiveness in addressing the examined market modeling problem, since market agents' actions are continuous and multi-dimensional.

In view of such challenges, *policy gradient* methods are preferred driven by their ability to handle continuous actions [32]. The main idea behind policy gradient method is to adjust the parameter θ in the direction of the *performance gradient* $\nabla_\theta \mathcal{J}(\pi_\theta)$, which is defined in the *policy gradient theorem* [30], [32]:

$$\begin{aligned}\nabla_\theta \mathcal{J}(\pi_\theta) &= \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a|s) Q^\pi(s, a) \mathrm{d}a\mathrm{d}s \\ &= \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \pi_\theta(a|s) \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)} Q^\pi(s, a) \mathrm{d}a\mathrm{d}s \\ &= \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) Q^\pi(s, a) \mathrm{d}a\mathrm{d}s \\ &= \mathbb{E}_{s\sim\rho^\pi, a\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q^\pi(s, a)] \end{aligned} \tag{18}$$

According to (18), to derive the policy gradient, one first needs to take samples of $a \sim \pi_\theta(a|s)$ and compute the estimated gradient as $\nabla_\theta \log \pi_\theta(a|s) Q^\pi(s, a)$. Moving a in the direction indicated by this gradient increases the log-probability of choosing that a proportionate to the associated action value function $Q^\pi(s, a)$. In this paper, we use the simple return R_t to estimate the value of $Q^\pi(s_t, a_t)$.

To this end, the deep policy gradient (DPG) network π_θ is a DNN, parameterized by θ, which takes as input a state s_t and performs the policy improvement task which updates the policy with respect to the estimated performance function $\mathcal{J}$ and outputs $\pi_\theta(a_t|s_t), \forall a_t \in \mathcal{A}$ which denotes the probability of take action a_t at state s_t. To update the DPG network, the policy gradients are placed at the network's output layer and then back-propagated through the network.

Concerning the way of the policy improvement, the approach employed in the Q-learning and DQN methods (Section III-C) involves a greedy maximization of the Q-value function, i.e., $\pi(s_{t+1}) = \mathrm{argmax}_{a_{t+1}} Q(s_{t+1}, a_{t+1})$. However, it is constructive to emphasize that in multi-dimensional continuous action spaces, greedy policy improvement becomes intractable as it necessitates maximizing the Q-value function globally. To address this challenge, the DPG network poses a more computationally friendly alternative which is to update agents' policy in the direction of the gradient of the performance function $\mathcal{J}$, rather than globally maximising the Q-value function.

During the learning process, state samples are generated as the agent sequentially interacts with the environment, suggesting that these samples are temporally correlated and does not meet the *independently and identically distributed* requirement of modern deep learning algorithms. To resolve this issue, an *experience reply buffer* $\mathcal{R}$ [31] is employed. This buffer is a cache of size $K_\mathcal{R}$ with a first-in-first-out queue rule which stores previous experiences (an experience is a transition tuple (s_t, a_t, r_t)). We then sample uniformly a minibatch of K trajectories of experiences $(s_1^{(k)}, a_1^{(k)}, r_1^{(k)}, ..., s_T^{(k)}, a_T^{(k)}, r_T^{(k)}), \forall k = 1, ..., K$ (a trajectory is an episode with T sequential time steps) to update DPG network parameters. Mixing recent with past experiences contributes to reducing the temporal correlations existing in the replayed experiences. Furthermore, the experience replay allows samples to be reused, and thereby enhances the sampling efficiency.

To this end, considering a sampled minibatch of K trajectories of experiences, the policy gradient can be

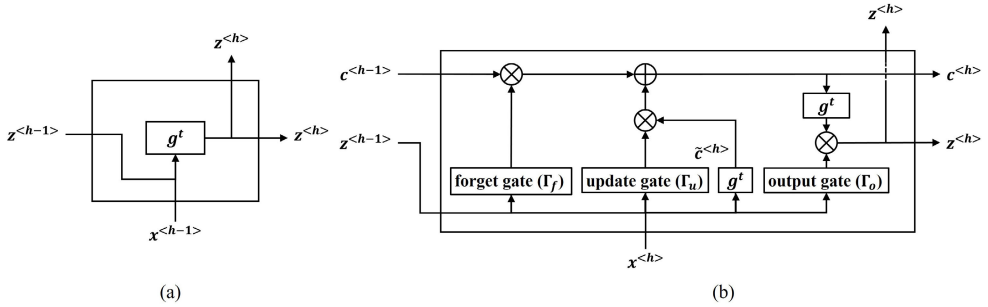


FIGURE 2. Structure of standard RNN cell and a LSTM cell.

expressed as:

$$\nabla_\theta \mathcal{J}(\pi_\theta) = \frac{1}{K}\sum_{k=1}^{K}\left[\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t^{(k)}|s_t^{(k)})\right)R_1^{(k)}\right] \quad (19)$$

where $R_1^{(k)} = \sum_{l=1}^{T}\gamma^{(l-1)}r_l^{(k)}$ is the total discount reward accumulated from the starting state of each trajectory. The following update is subsequently applied to update the weights of the DGP network, where α indictates the learning rates of the gradient decent algorithm:

$$\theta \leftarrow \theta + \alpha \cdot \nabla_\theta \mathcal{J}(\pi_\theta) \quad (20)$$

E. PROPOSED DPG-LSTM METHOD

Extracting discriminative features from raw state data is an imperative step toward improved learning performance of the proposed DPG network. In the examined market modeling problem, the perceived state of each GENCO i is a $3 \times N_H$-dimensional vector comprising the N_H generation dispatch of GENCO i, LMPs, and load forecasts. This raw data can be converted to multi-variate time-series data with N_H hours, each is characterized by three features, i.e. $g_{i,h}$, $\lambda_{(n:i\in I_n),h}$, and d_h. To effectively extract and interpret useful features of this time-series data, we propose a representation network as depicted in Fig. 1. The latter incorporates a Long Short-Term Memory (LSTM) network [34] which has gained significant research interest recently owning to its remarkable capability of capturing the long-range temporal dependencies of time-series data [34] compared to conventional feed-forward neural networks (i.e. DNN) and Recurrent Neural Networks (RNN). As such, LSTM networks have most recently received success in assorted power system/smart grid applications, such as electricity load [35] and electricity price forecasting [36].

The structure of a LSTM cell and a standard RNN cell is compared in Fig. 2. Given a temporal input sequence $[x^{<1:N_H>}]$ of length N_H, an RNN generates a sequence of output activation (or hidden) values $[z^{<1:N_H>}]$ by iterating the following recursive equation:

$$z^{<h>} = g^t(W_z[z^{<h>}, x^{<h>}] + b_z) \quad (21)$$

where $g^t(\cdot)$ denotes the hyperbolic tangent activation function, W_z is the matrix of weights and b_z is the vector of biases of appropriate sizes for the RNN cell.

LSTM network extends RNN with *memory cells* in order to store and output information, and thereby facilitating the learning of temporal dependencies for long duration of time. The idea of a LSTM network is based on a mechanism that defines the behavior of each individual memory cell, referring to as *gating*. The cell state of the LSTM network is denoted as $c^{<h>}$. The LSTM network then stores/removes information to/from the cell, governed by the operation of different gates. The equations expressing the operation of a LSTM cell are outlined below:

$$\tilde{c}^{<h>} = g^t(W_c[z^{<h-1>}, x^h] + b_c) \quad (22)$$
$$\Gamma_u = g^\sigma(W_u[z^{<h-1>}, x^h] + b_u) \quad (23)$$
$$\Gamma_f = g^\sigma(W_f[z^{<h-1>}, x^h] + b_f) \quad (24)$$
$$\Gamma_o = g^\sigma(W_o[z^{<h-1>}, x^h] + b_o) \quad (25)$$
$$c^{<h>} = \Gamma_u \odot \tilde{c}^{<h>} + \Gamma_f \odot \tilde{c}^{<h-1>} \quad (26)$$
$$y^{<h>} = z^{<h>} = \Gamma_o \odot g^t(c^{<h>}) \quad (27)$$

where W_c, W_u, W_f, W_o are the matrices of weights and b_c, b_u, b_f, b_o are vectors of bias of appropriate sizes for the LSTM cell. Equation (22) represents the input information. Equations (23)-(25) represent the operation of the update (or input), forget, and output gates where $g^\sigma(\cdot)$ denotes the sigmoid activation function. Equation (26) dictates the update of the memory cell state. The update Γ_i and forget Γ_f gates control, respectively, how much information to be written to the current cell state $c^{<h>}$ and how much information to be retained from the previous cell state $c^{<h-1>}$. Equation (27) indicates the output $y^{<h>}$ of the LSTM cell which in this case is the same as the output activation $z^{<h>}$ and is governed by the output gate Γ_o.

The overall workflow of the proposed DPG-LSTM method is illustrated in Fig. 1. The output layer of the LSTM network is a densely-connected layer with each neuron expresses the extracted features from raw data on generation dispatch, LMP, and load forecast. This layer is then connected to individual GENCO's DPG network (Section III-D). In stochastic continuous control RL problems, it is standard to represent the probability distribution of agent's action with a Normal

distribution $\mathcal{N}(\mu, \sigma^2)$, and predict the mean μ and the variance σ^2 of it with a DNN as the function approximator, referred to as a *Gaussian Policy*. In this context, the DPG network, parameterized by θ, takes the extracted feature vector as an input and outputs the Gaussian policy for each action dimension. As illustrated in Fig. 1, each GENCO i then selects its offering strategy by sampling from the obtained N_H Normal distributions according to:

$$\pi_{\theta_i}(a_{i,h}|s_i) \sim \mathcal{N}\left(\mu_h, \sigma_h^2\right), \quad \forall h = 1, \ldots, N_H. \tag{28}$$

F. DETERMINING OLIGOPOLISTIC MARKET EQUILIBRIUM WITH PROPOSED MA-DPG METHODOLOGY

The DPG-LSTM method enables each individual GENCO to learn its optimal offering strategy (Fig. 1). In order to determine the *Nash Equilibrium* (NE) under the participation of multiple strategic GENCOs, we propose a multi-agent DRL methodology, namely MA-DPG-LSTM (Fig. 1), which facilitates simultaneous learning of mutiple GENCOs' offering strategies and the analysis of the market outcome stemming from their interaction. In this case, each GENCO holds an experience reply buffer $\mathcal{R}_i$ which separately records the experiences of GENCO i gathered from its repeated interaction with the market clearing process (4)-(11). The policy gradient of GENCO i and the udapte of its DPG network can be expressed as (29) and (30), respectively.

$$\nabla_{\theta_i} \mathcal{J}(\pi_{\theta_i}) = \frac{1}{K} \sum_{k=1}^{K} \left[\left(\sum_{t=1}^{T} \nabla_{\theta_i} \log \pi_{\theta_i}(a_{i,t}^{(k)}|s_{i,t}^{(k)}) \right) R_{i,1}^{(k)} \right] \tag{29}$$

$$\theta_i \leftarrow \theta_i + \alpha \cdot \nabla_{\theta_i} \mathcal{J}(\pi_{\theta_i}) \tag{30}$$

The MA-DPG-LSTM method is outlined in Algorithm 1.

The relationship between multi-agent RL solution and NE is briefly discussed as follows. RL adopts differential learning mechanism to achieve Bellman optimality, which means RL is capable of learning the sub-game optimization substructure including NE [37]. However, in practice, it proves significantly challenging to gauge how close a collection of agents' strategies to a NE in large-scale games such as the market equilibrium problem investigated in this paper, due to the cost in training. As a result, researchers generally resort to convergence to control termination of the training process.

In the case where the proposed MA-DPG-LSTM method achieves convergence for all GENCOs (i.e. the offering strategies of all GENCOs remain constant (given some tolerance) with respect to the previous iteration), the diagonalization technique can be subsequently employed to verify whether the convergence state is a NE [13]. This method works by sequentially checking, for each GENCO i, whether its offering strategy (and profit) at convergence coincide with the respective solutions of its MPEC problem (Section I-A), holding the offering strategies of the rest of the GENCOs fixed and equal to their values at convergence. If the above holds for all GENCOs, then such convergence state corresponds by definition to a *pure strategy* NE of the oligopolistic market, since none of the GENCOs can increase their profits by unilaterally modifying their offering strategies.

Lastly, as discussed in the literature, existence and uniqueness of Nash equilibria are not generally guaranteed [11]–[17], [24], [25], [27]. However, an equilibrium has proven to be reached within a relatively small number of iterations in every examined case study (Section V). This finding, along with the fundamental contribution of this work on developing a MA-DRL based methodology to facilitate practical multi-period and multi-spatial equilibrium analysis in imperfect electricity markets, sets a detailed analysis of the determined equilibrium solutions out of the scope of this paper.

Algorithm 1 Proposed MA-DPG-LSTM Methodology

Initialize policy parameters θ_i for each GENCO i with random weights.
Initialize experience reply buffer $\mathcal{R}_i$ for each GENCO i, minibatch size K,
for episode $e = 1 : E$ **do**
  for GENCO $i = 1 : I$ **do** {in parallel}
    Selects random offer in its action space.
  end for
  The MO solves the market clearing problem (4)-(11) and announces the clearing outcome. The latter, along with the load forecast for day 1 is used as the initial state $s_{i,0}$ of GENCO i for the current episode.
  for trading day $t = 1 : T$ **do**
    for GENCO $i = 1 : I$ **do** {in parallel}
      Selects its offer $a_{i,t}$ using (28) according to its current policy π_{θ_i}.
    end for
    Based on the collected supply offers $(a_{i,t}, ..., a_{I,t})$, the MO solves problem (4)-(11) and broadcasts the market clearing outcome. This, along with the load forecast for day $t + 1$ serve as the new state $s_{i,t+1}$ for each GENCO i.
    for GENCO $i = 1 : I$ **do** {in parallel}
      Evaluate its profit / reward $r_{i,t}$ using (13).
      Stores, in its experience buffer $\mathcal{R}_i$, experience $(s_{i,t}, a_{i,t}, r_{i,t})$.
      Sample uniformly, from $\mathcal{R}_i$, a minibatch of K trajectories of accumulated experiences $(s_{i,1}^{(k)}, a_{i,1}^{(k)}, r_{i,1}^{(k)}, ..., s_{i,T}^{(k)}, a_{i,T}^{(k)}, r_{i,T}^{(k)})$.
      Update its DPG-LSTM network so as to update its policy π_{θ_i} using (29) and (30).
    end for
  end for
end for

IV. CASE STUDIES

A. TEST SYSTEM DATA AND IMPLEMENTATION

In this section, we validate the proposed MA-DPG-LSTM method in a test market with day-ahead horizon and hourly

resolution, operating over the IEEE Reliability Test System (RTS) [38] whose network topology is shown in Fig. 3.

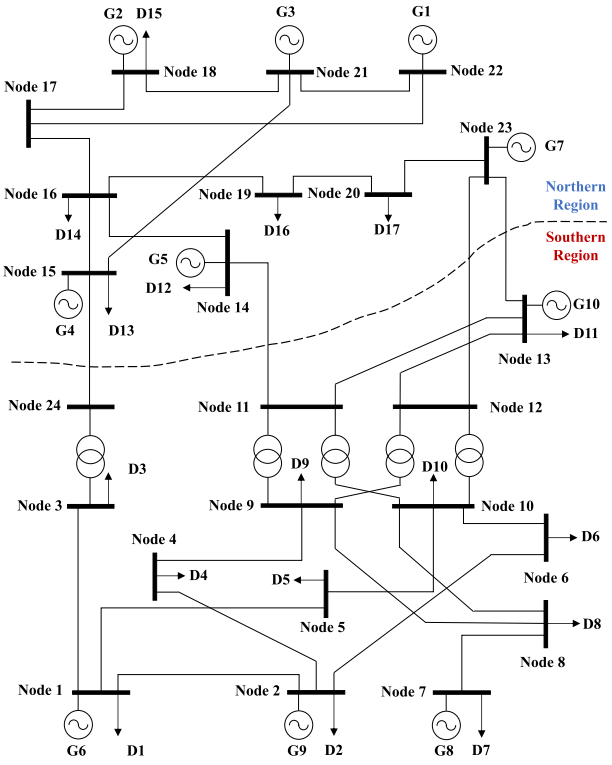


FIGURE 3. Network Topology of the IEEE RTS.

The market includes 10 GENCOs, with their location, marginal cost, maximum power output limit (assuming that one generation block is used), and ramp up / down limits provided in Table 1. The upper limits of the strategic offering variable of all GENCOs is assumed $\overline{o}_i = 2, \forall i$. The market also includes 17 demands, with their location and relative size (expressed as % of the total system demand and assuming that it remains identical for every time period) presented in Table 2. Fig. 4 presents the total demand profile of the system.

TABLE 1. Characteristics of GENCOs.

GENCO i	1	2	3	4	5	6	7	8	9	10
Node	22	18	21	15	14	1	23	7	2	13
λ_i^G (£/MWh)	2.6	3.8	4.2	4.8	5.0	5.5	7.2	8.2	9.0	10.0
$\overline{g}_i$ (MW)	160	140	130	120	100	70	50	60	30	30
R_i^U (MW)	100	80	80	60	50	50	40	40	30	30
R_i^D (MW)	100	80	80	60	50	50	40	40	30	30

To facilitate the analysis on the impact of network congestion, the RTS network is divided into two areas where nodes 1-13 and 24 correspond to the northern area while nodes 14-23 correspond to southern area. The northern area is characterized by cheaper generation and the largest demand centres are located in the southern area. This setting resembles a realistic situation for the Great Britain (GB) system [16], [17] where the northern / southern areas correspond to Scotland / England, respectively.

TABLE 2. Characteristics of demands.

Demand j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Node	1	2	3	4	5	6	7	8	9	10	13	14	15	16	18	19	20
Size (%)	4	4	13	3	3	7	7	7	11	11	13	2	2	2	7	3	3

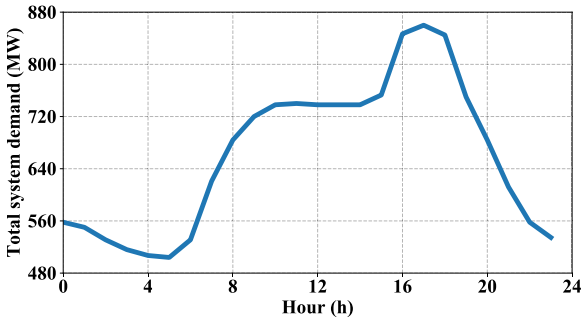


FIGURE 4. Total demand profile of the system.

It is worth mentioning that compared to the test systems examined in previous works [7]–[18], [21]–[29] (which represent state-of-the-art equilibrium programming papers in the literature), the examined test system in our paper involves the highest i) number of strategic GENCOs (10); ii) number of demand participants (17); and iii) number of time period (24). Taking into account all these points, the examined case study included present more complex cases with respect to this real-world electricity markets.

In order to validate the performance of the proposed DRL method, we compare it with Q-learning, DQN, and the original DPG methods which constitute the state-of-the-art RL methods in the power systems / smart grid literature. Their implementations are briefly discussed as follows.

1) Q-LEARNING

The RL problem must be calibrated in discrete state and action spaces (Section I-A) in order to apply Q-learning. In the examined market modeling problem, the states and actions correspond to the hourly LMPs and the hourly offering decisions respectively. We discretize the continuous states and actions in 100 integer values. Therefore, each GENCO i employs 24 look-up tables, each of size 100×100, to store and update the Q-values for state-action pairs at each hour h. Note that it is impractical to use a single look-up table of size $100^{24} \times 100^{24}$ to store the Q-values associated with different daily state-action pairs under the assumed discretization. Note also that although more state features (e.g. the generation dispatch) can be considered, it leads to exponentially increasing number of rows in the look-up table, rendering the problem intractable.

2) DQN

The DQN method makes use of a DNN as a function approximator that provides the Q-value estimate for each discrete action and, when acting, selects the action corresponding to the highest Q-value at a given state. In the examined market

modeling problem, the state is represented as a time-window of two adjacent hours, i.e., hour identifier h, dispatch of the GENCO i, LMP, and demand forecast at hours $h-1$ and h, resulting in 7 neurons in the input layer of the DNN. The continuous action space is discretized in the same fashion as in Q-learning, resulting in 100 neurons in the output layer of the DNN.

For the proposed DPG-LSTM method, the representation network features two LSTM layers with 8 and 16 neurons respectively. As shown in Fig. 1, a 384-dimension feature vector is extracted from the raw data comprising of generation dispatch, LMP, and demand forecast. This refined feature vector is subsequently fed into the input layer of the DPG network. The latter has two hidden layers with 128 and 64 neurons respectively and employ the rectified non-linearity (ReLU) [39] as activation function. The two output layers of the DPG network both have 24 neurons and encode the mean and standard deviation of the action, employing the sigmoid and softplus [39] as activation functions. The weights of the LSTM-based representation network and the DPG network are initialized with xavier initialization [39]. The Adam optimizer [39] is used for training the neural network weights with a learning rate $\alpha = 10^{-3}$. The discount factor γ is set to be 0.95. We train with a minibatch size of 32 and an experience replay buffer of size 128.

The examined RL methods have been implemented in Python with Tensorflow 1.12.0 [40]. The market clearing algorithm (4)-(11) and all the examined game-theoretic approaches (MPEC, EPEC, DIAG, and MILP) have been implemented using Xpress Optimizer Python interface [41]. The case studies have been carried out on a computer with a 6-core 3.50 GHz Intel(R) Xeon(R) E5-1650 v3 processor and 32 GB of RAM.

B. COMPARISON OF PERFORMANCE OF RL AND MPEC METHODS: SINGLE GENCO'S STRATEGIC OFFERING PROBLEM

The aim of this section lies in comparing the performance of different RL methods in terms of the quality (i.e. the profitability) of the learned offering strategy. In this context, we focus on a single GENCO's optimal strategic offering problem. In the examined case studies, this corresponds to GENCO 7 of Table 1 while the rest of the GENCOs are assumed to be price-takers (i.e. $o_{i,t} = 1, \forall i \in I \backslash \{7\}, \forall t$). For the sake of comparison clarity, this section considers a case where the network capacity limits are neglected. We randomly generate 10 different seeds, and for each seed each RL method is trained for 100 episodes, where an episode is composed of 20 time steps (Algorithm 1).

Fig. 5 illustrates the episodic average reward (i.e. the profit calculated using (13)) with 10 different random seeds for each of four examined RL methods and for the benchmark method where GENCO 7 directly optimizes its offering strategy through the state-of-the-art MPEC method (Section I-A). The lines and the shaded area depict the mean and standard deviation of the average reward over the 10 different random seeds. As shown in Fig. 5, the average profit is comparatively low during the initial phase of learning, suggesting that GENCO 7 is accumulating more experiences by randomly exploring different actions. As the learning continues and more experiences being accumulated, the average reward turns positive and keeps incr easing and eventually reaches convergence for all four RL models. This is reflected in the stabilized average reward as well as the decreased standard deviation as the learning approaches to the end (Fig. 5). The training of DPG-LSTM and DPG initially exhibit relatively larger variability compared to Q-learning and DQN. This is because exploring in multi-dimensional continuous action space (i.e. DPG and DPG-LSTM) is more challenging than in discrete action space (i.e. Q-learning and DQN). Nevertheless, as the learning process continues, DPG-LSTM significantly outperforms the other two methods, obtaining the highest average profit and exhibiting the smallest standard deviation at convergence. In relative terms, DPG-LSTM achieves 43.35% / 15.48% higher average profit and 63.87% / 49.31% lower standard deviation over Q-learning / DQN respectively. Furthermore, the DPG-LSTM method approximates very closely the profit obtained by the MPEC method, which in this case provides the benchmark solution (the difference between the profits obtained by the two methods is 0.90%). Finally, the DPG-LSTM method outperforms the original DPG method without the LSTM-based representation network (Section III-D), achieving 4.43% higher profit.

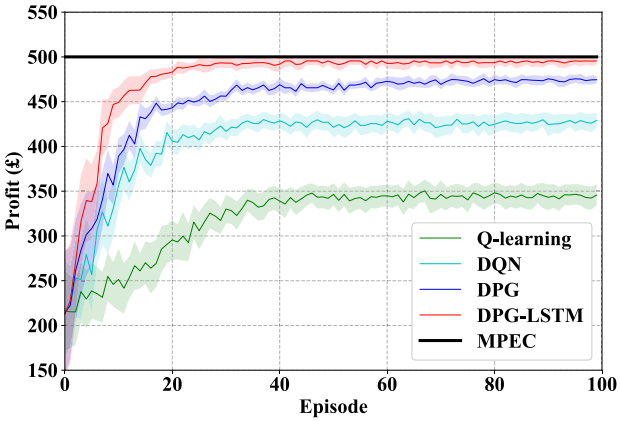


FIGURE 5. Episodic average reward over 10 different random seeds for the examined RL and MPEC methods.

The superior performance of DPG-LSTM can be explained by i) its ability to model multi-dimensional continuous state space and to extract discriminative features from time-series state vector using the proposed LSTM-based representation network, in contrast to discrete scalar states employed in Q-learning, enabling GENCO to receive accurate feedback regarding the impact of its offering strategies on the multi-period market clearing outcome and ii) its ability to model multi-dimensional continuous action space enabled by the proposed DPG network, in contrast to the naïve discretization approach employed in Q-learning and DQN, enabling GENCO 7 to preserve more accurate information regarding the entire action space.

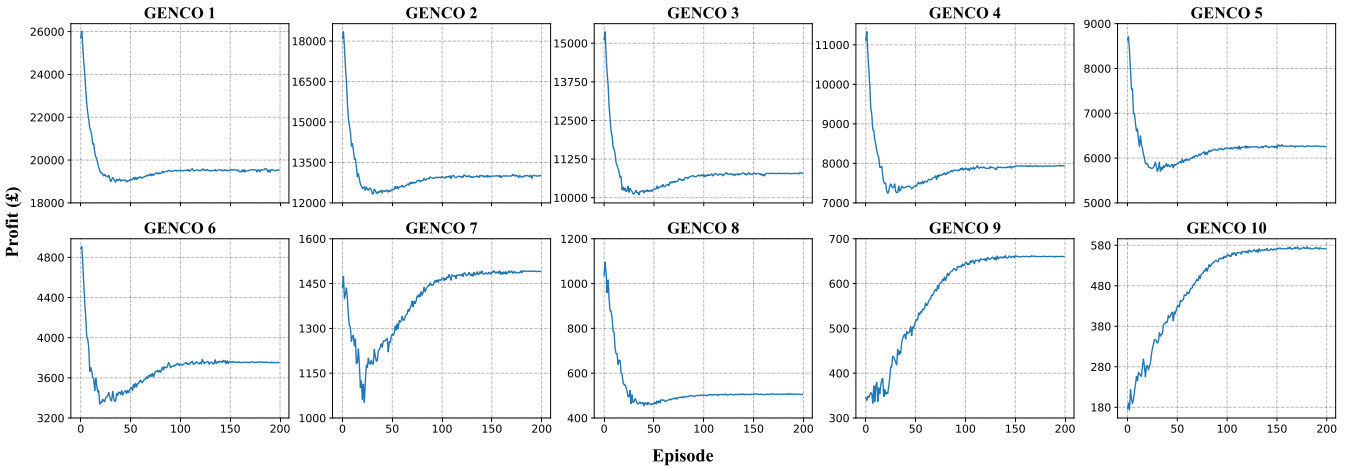


FIGURE 6. Episodic average profit for each of the 10 GENCOs in the oligopolistic market case without network congestion.

C. MULTI-PERIOD EQUILIBRIUM ANALYSIS: UNCONGESTED NETWORK

This section concerns the analysis of the multi-period market equilibrium where the network capacity limits are neglected and therefore the network is not congested. An imperfect, oligopolistic market is considered, where the offering strategies of the GENCOs are determined based on the proposed MA-DPG-LSTM methodology. Fig. 6 illustrates the episodic average profit for each of the 10 GENCOs in the oligopolistic market case. It can be observed that the average profits for all 10 GENCOs reach stabilization in around 150 episodes. The procedure of verification of NE presented in Section III-F) is conducted and it is confirmed that the obtained convergence state is indeed an NE since no GENCO sees any reason to deviates its decision given the rest of the GENCOs do not deviate from their decisions.

Fig. 7 illustrates the evolution of episodic average market prices in the oligopolistic market case. The intense competition among GENCOs contributes to the decreasing of market prices during the off-peak periods, where the available generation capacity is considerably larger than the demand. However, during the peak period, driven by the increasing slope of the GENCOs' offering curves at higher demand levels and the higher need to utilize available generation capacity in the system, the market prices are increased due to the GENCOs learn to exercise market power. Therefore, peak periods are deemed the most critical ones concerning the exercise of market power by strategic GENCOs [16], [17]. Table 3 presents the comparison of offering strategies and generation dispatch of GENCOs 8 and 10 at hours 17-19 in the oligopolistic market case. In the equilibrium, the most costly unit GENCO 10 selects lower offering strategies than GENCO 8 to sell more energy to the market (it is fully dispatched at hours 17-19) whilst GENCO 8 exercises market power to its largest extent and becomes the marginal unit and sets the market prices at 16.4 £/MWh at hours 17-19. These findings demonstrate the effectiveness of the proposed MA-DPG-LSTM method in learning the optimal offering strategies at different hours for GENCOs of different merit orders in the oligopolistic equilibrium of the market.

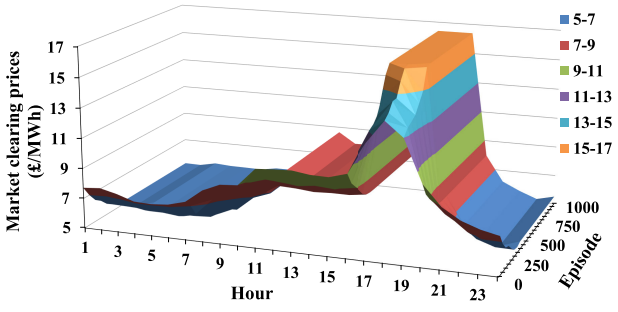


FIGURE 7. Evolution of the episodic average market prices in the oligopolistic market case.

TABLE 3. Offering strategies and generation dispatches of GENCOs 8 and 10 at hours 17-19 in the oligopolistic market case without network congestion.

$GENCO_i$	Offering strategy $o_{i,h}$			Dispatch $g_{i,h}$ (MW)		
	Hour 17	Hour 18	Hour 19	Hour 17	Hour 18	Hour 19
8	2.00	2.00	2.00	17	30	15
10	1.01	1.02	1.11	30	30	30

D. MULTI-PERIOD AND MULTI-SPATIAL EQUILIBRIUM ANALYSIS: CONGESTED NETWORK

This section presents the analysis of the multi-period and multi-spatial market equilibrium where the impact of network congestion is accounted for. Fig. 8 illustrates the episodic average profit for each of the 10 GENCOs in the oligopolistic market case. Similar to the trend observed in Fig. 6, the average profit for all 10 GENCOs reaches stabilization in around 175 episodes. The verification of NE is analogously carried out which confirms the convergence state is indeed an NE.

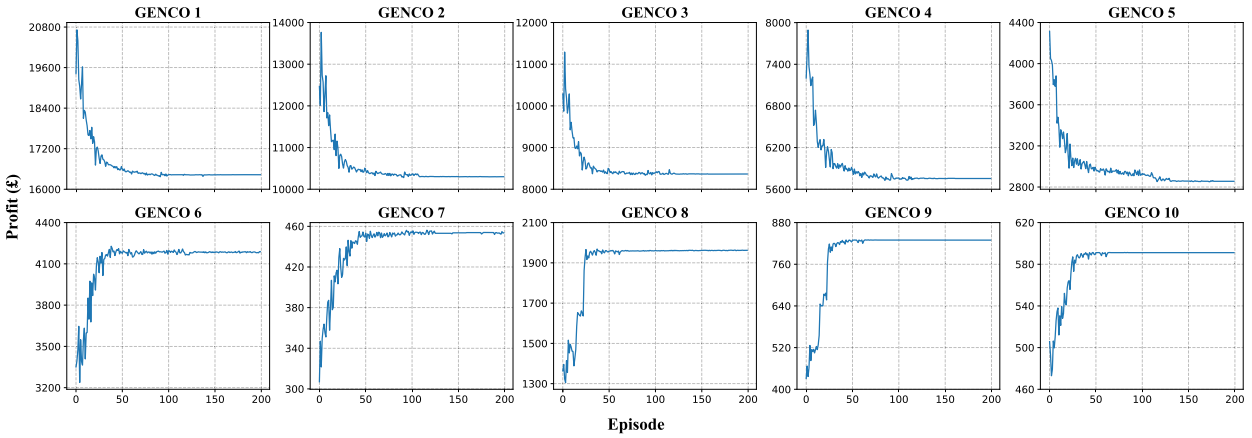


FIGURE 8. Episodic average profit for each of the 10 GENCOs in the oligopolistic market case with network congestion.

TABLE 4. Offering strategies and generation dispatches of GENCOs 5, 7, 8, and 9 at hours 17-19 for cases U and C in the oligopolistic market case.

$GENCO_i$	Case	Offering strategy $o_{i,h}$					Dispatch $g_{i,h}$ (MW)				
		Hour 16	Hour 17	Hour 18	Hour 19	Hour 20	Hour 16	Hour 17	Hour 18	Hour 19	Hour 20
5	**U**	1.25	1.20	1.62	1.41	1.25	100	100	100	100	100
	C	1.55	2.00	2.00	2.00	1.55	90	90	91	90	90
7	**U**	1.10	1.10	1.30	1.27	1.10	33	50	50	50	30
	C	1.14	2.00	2.00	2.00	1.17	31	31	30	30	30
8	**U**	1.08	2.00	2.00	2.00	1.07	0	17	30	15	0
	C	1.19	1.99	1.38	1.57	1.17	12	60	60	60	10
10	**U**	1.11	1.01	1.02	1.11	1.11	0	30	30	30	0
	C	1.15	2.00	2.00	2.00	1.27	0	16	29	14	0

In order to analyze the impact of network congestion, the following two cases are examined:

U: a case of oligopolistic market, where the network is not congested, which is identical in Section IV-D;

C: a case of oligopolistic market, where the network capacity limits are taken into account, in this case the lines (11-14) and (13-23) connecting northern and southern areas get congested during some peak periods, reflecting a realistic condition where network corridors connecting northern and southern areas are congested due to the transmission of northern cheaper generation to southern large demand centres (Fig. 3, Tables 1 and 2).

Table 4 presents the offering strategies and generation dispatches of GENCOs 5, 7, 8, and 10 at the critical congestion periods (hours 16-20) for cases U and C in the oligopolistic market case. Fig. 9 illustrates the 24-hour LMPs at nodes 11, 13, 14 and 23 for cases U and C in the oligopolistic market case. When the network is congested, the power flow from the northern to southern area is limited, which reduces / increases the dispatch of certain GENCOs located in the northern / southern area. In the oligopolistic equilibrium of case C, in the northern area, GENCOs 5 and 7 choose higher offering strategies (than in case U) at the congested hours, as the network capacity limit is restricting them from selling more energy to the southern area.[1] In this case, GENCOs 5 and 7

[1]Recall that the key in selecting the optimal offering strategy is to achieve an advantageous trade-off between increasing the market prices and increasing quantity sold to the market (Section II-B).

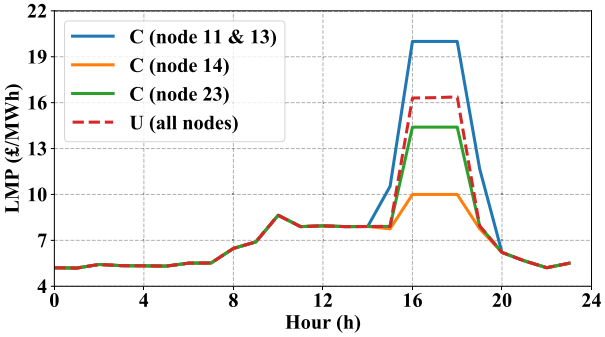


FIGURE 9. LMP at nodes 11, 13, 14, and 23 for cases U and C in the oligopolistic market case.

constitute the marginal units and determine the LMPs at node 14 and 23 in the northern area, respectively. In the southern area, driven by a combined effect of i) GENCO 10 (which is the most expensive unit (Table 1)) chooses a higher offering strategy and ii) the locational decoupling effect of congestion, the LMPs at the congested hours in the southern area are raised significantly but the energy sold by GENCO 10 is reduced. Given the latter and the reduced import from the northern area, GENCO 8 with lower marginal cost and higher capacity undercuts GENCO 10 and produces more energy in order to meet the southern demand. GENCO 8's profitability is consequently enhanced by benefiting from the high LMPs set by GENCO 10 during the congested hours as well as selling more energy to the market.

As shown in Fig. 9, when the network is not congested, the LMPs are identical at every node of the transmission network, while congestion in lines (11,14) and (13,23) yields locational price differential between the two areas. More specifically, during periods of congestion (hour 16-20), southern area (nodes 11 and 13) -featuring more costly generation and higher demand- exhibits a higher price than the one observed in the case U, while northern area (nodes 14 and 23) -featuring less costly generation and lower demand- exhibits a lower price than the one observed in case U. Table 5 presents the profits of GENCOs for cases U and C in the oligopolistic market case. As can be observed, network congestion creates a more favourable economic setting (i.e. higher profit) for GENCOs in southern area and a less favourable setting (i.e. lower profit) for GENCOs in northern area, as indicated by the profit increments of GENCOs in Table 5.

TABLE 5. Profits of GENCOs for cases U and C in the oligopolistic market case.

$GENCO_i$	Area	pro_i^U (£)	pro_i^C (£)	Profit Increment (£)
1	North	19452	16350	-3102
2	North	12988	10267	-2721
3	North	10812	8320	-2493
4	North	7909	5676	-2233
5	North	6175	2880	-3295
6	South	3692	4178	487
7	North	1495	452	-1043
8	South	503	1961	1458
9	South	660	829	169
10	South	570	591	21

These findings demonstrate the effectiveness of the proposed MA-DPG-LSTM method in learning the optimal offering strategies at different hours for GENCOs of different merit orders in the oligopolistic equilibrium of the market.

E. COMPUTATIONAL PERFORMANCE COMPARISON AGAINST CONVENTIONAL EQUILIBRIUM PROGRAMMING METHODS

The aim of this section lies in comparing the computational performance of the proposed MA-DPG-LSTM method against three conventional equilibrium programming models including the EPEC, DIAG, and MILP approaches (Section I-A). Table 6 summerised the computational performance of these approaches by presenting the total computational time required by the examined four methods to find a NE for cases U and C. As shown in Algorithm 1 (Section III-F), in each episode, each GENCO trains its own DPG-LSTM model by interacting with the market clearing process. The training process of each DPG-LSTM is implemented in a paralleled fashion. If a convergence state is observed and passes the NE verification test, the total computational time required for reaching the convergence state (indicated by the average profit for all 10 GENCOs reach stabilization) is then recorded.

TABLE 6. Computational time (minutes) for finding a NE in each of the examined cases. (*: No solution found after 24 hours of simulation).

Cases	EPEC	DIAG	MILP	MA-DPG-LSTM
U	66	82	*	13
C	130	144	*	15

It can be observed that the proposed MA-DPG-LSTM method finds the NE for both cases U and C in approximately 13 and 15 minutes, respectively. This is while the MILP approach fail to identify any NE after 24 hour of simulation. Moreover, although EPEC and DIAG approaches can eventually locate a NE, their computational intensity is much higher than the proposed MA-DPG-LSTM method (Table 6). The reason behind the unsatisfactory performance of the MILP approach lies in the vast number of the binary variables included in the model [18]. Also, the convergence of branch and bound solvers highly depends on tuning the disjunctive (or big-M) parameters introduced for pursing linearity The inherent non-convexities and non-linearities presented in the EPEC formulation -driven by a large number of complementarity constraints and the mixed-integer linearization of the bilinear terms in the model- renders them very hard and expensive to solve. For the DIAG approach, at each iteration, multiple MPEC problems (which are non-smooth and non-convex in nature) need to be solved, making it very computational demanding as well. Furthermore, all the aforementioned computational complexities are aggravated in the examined multi-period, network-constrained market modeling problem, rending these approaches less useful for finding a NE. Lastly, although none of the examined methods can theoretically guarantee their solution existence or convergence to a NE [10], [11], [27], [42], case studies demonstrate that MA-DPG-LSTM exhibits superior computational performance in successfully and efficiently identifying a NE in a multi-period and network-constrained electricity market.

V. CONCLUSION AND FUTURE WORK

Existing literature largely resort to conventional game-theoretic approaches for modeling and analyzing imperfect electricity markets. However, such approaches exhibit severe modeling and computational complexities and are thus very hard and computationally expensive to solve. In addition, they rely on complete knowledge of the techno-economical characteristics and the strategies of the market players as well as the computational algorithm of the market clearing process, which piratically constitutes a very constraining assumption. Furthermore, such approaches overlook the accumulated experiences of learning from GENCOs' daily repeated interactions with the market clearing.

In view of these limitations, this paper has proposed a novel MA-DRL based methodology, combining multi-agent intelligence and a DPG-LSTM method, to expedite

practical multi-period and multi-spatial equilibrium analysis. In contrast with state-of-the-art RL methods (Q-learning and DQN), this approach conforms to the nature of the examined problem in multi-dimensional continuous state and action spaces, enabling GENCOs to receive accurate feedback regarding the impact of their offering strategies on the market clearing outcome, and devise more profitable bidding decisions by exploring the entire action domain. Furthermore, the proposed LSTM-based representation network further improves GENCOs' profitability driven by its ability to extract high-dimensional discriminative features from raw data on the market condition and clearing outcome.

Case studies on a test market with day-ahead horizon and hourly resolution and operated over the IEEE RTS system have validated the effectiveness of the proposed methodology. Regarding the single GENCO's optimal offering strategy problem, the proposed DPG-LSTM method promises a substantially higher profit than state-of-the-art RL methods (Q-learning, DQN, and DPG) and approximates very closely the profit obtained by the state-of-the-art MPEC method. Concerning the equilibrium programming problem, the proposed MA-DPG-LSTM method outperforms state-of-the-art equilibrium programming models (EPEC, DIAG, and MILP) in efficiently discovering an imperfect market equilibrium. Quantitative economic analysis has been carried out on the obtained equilibrium. In the case without network congestion, results have demonstrated MA-DPG-LSTM is able to learn the optimal offering strategies at different hours for GENCOs of different merit orders. In cases with network congestion, GENCOs located in the higher-priced area learn to evolve their strategies by exploiting the price differential effect created by the congestion, attaining higher profits albeit at the expense of the profitability of the GENCOs located in the lower-priced area.

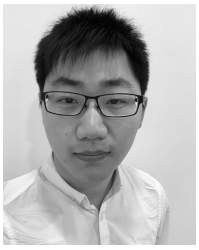
Conventional equilibrium programming models in the existing literature [10]–[17] neglect the complex unit commitment constraints of the generation units, due to their intrinsic inability to deal with binary decision variables in the LL problem of the strategic GENCOs' bi-level optimization problems. However, these complex operating properties may affect the market clearing outcome and consequently the strategic decisions of the market players. In contrast, under the proposed MA-DPG-LSTM method, the bi-level optimization problem is not converted to a single-level, closed-form MPEC. Instead, it is solved in a recursive fashion where strategic GENCOs gradually learn their optimal offering strategies from repeated interactions with the market clearing process. It therefore avoids the derivation of the equivalent KKT optimality conditions of the LL problem and is capable of addressing the aforementioned challenge of incorporating non-convex operating characteristics into the market clearing process. Future work aims at extending the proposed MA-DPG-LSTM method to investigate the strategic behaviour of GENCOs as well as the market outcomes stemming from their interactions.

REFERENCES

[1] D. S. Kirschen and G. Strbac, *Fundamentals of Power System Economics*, 2nd ed. West Sussex, U.K.: Wiley, 2018.
[2] D. R. Biggar and M. R. Hesamzadeh, *The Economics of Electricity Markets*. Hoboken, NJ, USA: Wiley, 2014.
[3] A. G. Bakirtzis, N. P. Ziogos, A. C. Tellidou, and G. A. Bakirtzis, "Electricity producer offering strategies in day-ahead energy market with step-wise offers," *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 1804–1818, Nov. 2007.
[4] C. Ruiz and A. J. Conejo, "Pool strategy of a producer with endogenous formation of locational marginal prices," *IEEE Trans. Power Syst.*, vol. 24, no. 4, pp. 1855–1866, Nov. 2009.
[5] E. G. Kardakos, C. K. Simoglou, and A. G. Bakirtzis, "Optimal bidding strategy in transmission-constrained electricity markets," *Electr. Power Syst. Res.*, vol. 109, pp. 141–149, Apr. 2014.
[6] L. Xu, R. Baldick, and Y. Sutjandra, "Bidding into electricity markets: A transmission-constrained residual demand derivative approach," *IEEE Trans. Power Syst.*, vol. 26, no. 3, pp. 1380–1388, Aug. 2011.
[7] G. Federico and D. Rahman, "Bidding in an electricity pay-as-bid auction," *J. Regulatory Econ.*, vol. 24, no. 2, pp. 175–211, Sep. 2003.
[8] D. Acemoglu, A. Kakhbod, and A. Ozdaglar, "Competition in electricity markets with renewable energy sources," *Energy J.*, vol. 38, pp. 137–155, Sep. 2017.
[9] E. Bompard, W. Lu, R. Napoli, and X. Jiang, "A supply function model for representing the strategic bidding of the producers in constrained electricity markets," *Int. J. Elect. Power Energy Syst.*, vol. 32, no. 6, pp. 678–687, Jul. 2010.
[10] W. Xian, L. Yuzeng, and Z. Shaohua, "Oligopolistic equilibrium analysis for electricity markets: A nonlinear complementarity approach," *IEEE Trans. Power Syst.*, vol. 19, no. 3, pp. 1348–1355, Aug. 2004.
[11] J. Yao, I. Adler, and S. S. Oren, "Modeling and computing two-settlement oligopolistic equilibrium in a congested electricity network," *Oper. Res.*, vol. 56, no. 1, pp. 34–47, Feb. 2008.
[12] C. Ruiz, A. J. Conejo, and Y. Smeers, "Equilibria in an oligopolistic electricity pool with stepwise offer curves," *IEEE Trans. Power Syst.*, vol. 27, no. 2, pp. 752–761, May 2012.
[13] A. Shahmohammadi, R. Sioshansi, A. J. Conejo, and S. Afsharnia, "Market equilibria and interactions between strategic generation, wind, and storage," *Appl. Energy*, vol. 220, pp. 876–892, Jun. 2018.
[14] C. A. Berry, B. F. Hobbs, W. A. Meroney, R. P. O'Neill, and W. R. Stewart, Jr., "Analyzing strategic bidding behavior in transmission networks," *Utilities Policy*, vol. 8, no. 3, pp. 139–158, Jan. 1999.
[15] J. D. Weber and T. J. Overbye, "An individual welfare maximization algorithm for electricity markets," *IEEE Trans. Power Syst.*, vol. 17, no. 3, pp. 590–596, Aug. 2002.
[16] Y. Ye, D. Papadaskalopoulos, and G. Strbac, "Investigating the ability of demand shifting to mitigate electricity producers' market power," *IEEE Trans. Power Syst.*, vol. 33, no. 4, pp. 3800–3811, Jul. 2018.
[17] Y. Ye, D. Papadaskalopoulos, R. Moreira, and G. Strbac, "Investigating the impacts of price-taking and price-making energy storage in electricity markets through an equilibrium programming model," *IET Gener., Transmiss. Distrib.*, vol. 13, no. 2, pp. 305–315, Jan. 2018.
[18] M. R. Hesamzadeh and D. R. Biggar, "Computation of extremal-Nash equilibria in a wholesale power market using a single-stage MILP," *IEEE Trans. Power Syst.*, vol. 27, no. 3, pp. 1706–1707, Aug. 2012.
[19] D. Pozo, E. Sauma, and J. Contreras, "Basic theoretical foundations and insights on bilevel models and their applications to power systems," *Ann. Oper. Res.*, vol. 254, nos. 1–2, pp. 303–334, Jul. 2017.
[20] A. Weidlich and D. Veit, "A critical survey of agent-based wholesale electricity market models," *Energy Econ.*, vol. 30, no. 4, pp. 1728–1759, Jul. 2008.
[21] G. Xiong, T. Hashiyama, and S. Okuma, "An electricity supplier bidding strategy through Q-learning," in *Proc. IEEE Power Eng. Soc. Summer Meeting*, Chicago, IL, USA, vol. 3, Jul. 2002, pp. 1516–1521.
[22] R. Ragupathi and T. K. Das, "A stochastic game approach for modeling wholesale energy bidding in deregulated power markets," *IEEE Trans. Power Syst.*, vol. 19, no. 2, pp. 849–856, May 2004.
[23] M. B. Naghibi-Sistani, M. R. Akbarzadeh-Tootoonchi, M. H. J.-D. Bayaz, and H. Rajabi-Mashhadi, "Application of Q-learning with temperature variation for bidding strategies in market based power systems," *Energy Convers. Manage.*, vol. 47, nos. 11–12, pp. 1529–1538, Jul. 2006.
[24] T. Krause, E. V. Beck, R. Cherkaoui, A. Germond, and D. Ernst, "A comparison of Nash equilibria analysis and agent-based modelling for power markets," *Int. J. Elect. Power Energy Syst.*, vol. 28, no. 9, pp. 599–607, Nov. 2006.

[25] V. Nanduri and T. K. Das, "A reinforcement learning model to assess market power under auction-based energy pricing," *IEEE Trans. Power Syst.*, vol. 22, no. 1, pp. 85–95, Feb. 2007.
[26] M. Rahimiyan and H. R. Mashhadi, "Supplier's optimal bidding strategy in electricity pay-as-bid auction: Comparison of the Q-learning and a model-based approach," *Electr. Power Syst. Res.*, vol. 78, no. 1, pp. 165–175, Jan. 2008.
[27] V. Nanduri and T. K. Das, "A reinforcement learning algorithm for obtaining the Nash equilibrium of multi-player matrix games," *IIE Trans.*, vol. 41, no. 2, pp. 158–167, Feb. 2009.
[28] N.-P. Yu, C.-C. Liu, and J. Price, "Evaluation of market rules using a multi-agent system method," *IEEE Trans. Power Syst.*, vol. 25, no. 1, pp. 470–479, Feb. 2010.
[29] H. Kebriaei, A. Tajeddini, and N. Rashedi, "Markov game approach for multi-agent competitive bidding strategies in electricity market," *IET Gener., Transmiss. Distrib.*, vol. 10, no. 15, pp. 3756–3763, Nov. 2016.
[30] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[31] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.
[32] R. S. Sutton, D. A. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, Denver, CO, USA, 2000, pp. 1057–1063.
[33] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, May 1992.
[34] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.
[35] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, Jan. 2019.
[36] L. Peng, S. Liu, R. Liu, and L. Wang, "Effective long short-term memory with differential evolution algorithm for electricity price prediction," *Energy*, vol. 162, pp. 1301–1314, Nov. 2018.
[37] A. Nowé, P. Vrancx, and Y.-M. De Hauwere, "Game theory and multi-agent reinforcement learning," *Reinforcement Learning* (Adaptation, Learning, and Optimization), vol. 12. Jan. 2012, pp. 441–470.
[38] C. Grigg, P. Wong, P. Albrecht, R. Allan, M. Bhavaraju, R. Billinton, Q. Chen, C. Fong, S. Haddad, S. Kuruganty, W. Li, R. Mukerji, D. Patton, N. Rau, D. Reppen, A. Schneider, M. Shahidehpour, and C. Singh, "The IEEE reliability test system-1996. A report prepared by the reliability test system task force of the application of probability methods subcommittee," *IEEE Trans. Power Syst.*, vol. 14, no. 3, pp. 1010–1020, Aug. 1999.
[39] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[40] M. Abadi. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: https://www.tensorflow.org/
[41] (2017). *Xpress Optimizer Python Interface*. [Online]. Available: https://www.msi-jp.com/xpress/learning/square/01-python-interface.pdf
[42] S. A. Gabriel, A. J. Conejo, J. D. Fuller, B. F. Hobbs, and C. Ruiz, *Complementarity Modeling in Energy Markets*, vol. 180. New York, NY, USA: Springer, 2012.

YUJIAN YE (M'12) received the B.Eng. degree (Hons.) in electrical and electronic engineering from Northumbria University, Newcastle upon Tyne, U.K., in 2011, and the M.Sc. degree in control systems and the Ph.D. degree in electrical engineering research from Imperial College London, London, U.K., in 2012 and 2016, respectively. He is currently a Machine Learning Scientist with Feth.AI, Cambridge, U.K., and a Research Associate with Imperial College London. His current research interests include development and application of novel game-theoretic and agent-based modeling methodologies in energy markets and smart energy systems as well as distributed control approaches for the coordination of operation and planning decisions in power systems.



DAWEI QIU (S'19) received the B.Eng. degree (Hons.) in electrical and electronic engineering from Northumbria University, Newcastle upon Tyne, U.K., in 2014, and the M.Sc. degree in power system engineering from University College London, London, U.K., in 2015. He is currently pursuing the Ph.D. degree with Imperial College London, London. His current research interests include developing game-theoretic and reinforcement learning approaches in wholesale and retail electricity market modeling problems.

JING LI (S'18) received the B.S. degree in automation from the Beijing Institute of Technology, Beijing, China, in 2016, and the M.Sc. degree in control systems from Imperial College London, London, U.K., in 2017, where she is currently pursuing the Ph.D. degree. Her current research interests include development of distributed control and game-theoretic approaches in local energy market design.

GORAN STRBAC (M'95) received the B.Sc. degree from the University of Novi Sad, Novi Sad, Serbia, in 1984, and the M.Sc. and Ph.D. degrees from the University of Belgrade, Belgrade, Serbia, in 1989 and 1994, respectively, all in electrical engineering. He was with the University of Manchester Institute of Science and Technology (UMIST), Manchester, U.K., for 11 years. He has been a Professor of electrical energy systems with Imperial College London, London, U.K., since 2005. He led the development of novel advanced analysis approaches and methodologies that have been extensively used to inform industry, governments, and regulatory bodies about the role and value of emerging new technologies and systems in supporting cost-effective evolution to smart low-carbon future. His research interests include modeling and optimization of economics and security of energy system operation and investment, energy infrastructure reliability, and future energy markets, including integration of emerging technologies in supporting cost-effective evolution to smart low-carbon energy future. He is also a member of the Steering Committee of the Smart Grids European Technology Platform, the Co-Chair of EU WG on Sustainable Districts and Built Environment of Smart Cities, and the Director of the U.K. Centre for Grid Scale Energy Storage. He participate in working groups and committees within CIGRE, CIRED IET, the IEEE, and IEA.