IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Distributed Optimization for Distribution Grids with Stochastic DER using Multi-Agent Deep Reinforcement Learning

MOHAMMED AL-SAFFAR[1], (Member, IEEE), AND PETR MUSILEK.[1,2], (Senior Member, IEEE)
[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada
[2]Department of Cybernetics, University of Hradec Králové, Hradec Králové, Czech Republic

Corresponding author: Mohammed Al-Saffar (e-mail: malsaffa@ualberta.ca).

This work was supported in part by the Canada First Research Excellence Fund (CFREF) under the Future Energy Systems research initiative at the University of Alberta, and by the Natural Science and Engineering Research Council (NSERC) of Canada under grant RGPIN-2017-05866."

ABSTRACT This article develops a special decomposition methodology for the traditional optimal power flow which facilitates optimal integration of stochastic distributed energy resources in power distribution systems. The resulting distributed optimal power flow algorithm reduces the computational complexity of the conventional linear programming approach while avoiding the challenges associated with the stochastic nature of the energy resources and loads. It does so using machine learning algorithms employed for two crucial tasks. First, two proposed algorithms, Dynamic Distributed Multi-Microgrid and Monte Carlo Tree Search based Reinforcement Learning, constitute dynamic microgrids of network nodes to confirm the electric power transaction optimality. Second, the optimal distributed energy resources are obtained by the proposed deep reinforcement learning method named Multi Leader-Follower Actors under Centralized Critic. It accelerates conventional linear programming approach by considering a reduced set of resources and their constraints. The proposed method is demonstrated through a real-time balancing electricity market constructed over the IEEE 123-bus system and enhanced using price signals based on distribution locational marginal prices. This application clearly shows the ability of the new approach to effectively coordinate multiple distribution system entities while maintaining system security constraints.

INDEX TERMS Distributed architecture, distributed optimization, Monte Carlo tree search, multi-agent deep reinforcement learning, optimal power flow.

I. INTRODUCTION

OPTIMAL power flow (OPF) is an essential tool for managing energy in electric power systems. It seeks the least cost operation of a power system by dispatching generation for given power demand while satisfying the system constraints. The changing nature of modern power grids brings new entities into electric power markets. They include owners of distributed energy resources (DERs), and even so called prosumers - individual customers equipped with self-owned DER units. The new market participants are interested in autonomous maximization of their profits. Therefore, they can be considered independent entities of the system [1]. However, a decision made by a single entity may affect the decisions of the remaining entities that are physically interconnected in the same system.

As power distribution systems are becoming more and more dispersed, they may require additional generation capacity and new line assets to supply the peak demand. The network participants may need to cooperate with each other to achieve reliable and effective operation of the network without changing the system infrastructure. The incremental dispersion of new network entities will also affect the electric power markets. In this new scenario, the interactions between two independent bilateral power transactions in the network need to be checked and optimized using OPF. However, the conventional centralized OPF method poses a number of problems [2]. To avoid these issues and provide the power industry with tools to support highly efficient system operation, distributed optimization architectures are required. Such architectures can capture all physical realities of a dispersed network and alleviate a centralized optimization agent from tremendous amount of computing.

A. RELATED WORK

Recent literature presents several approaches to distributed economic dispatch [3], [4]. They resolve the randomness of DER units and loads in microgrids through the use of Markov decision processes (MDPs). Distributed model predictive control (MPC) for stochastic dispatch optimization in microgrids have been proposed by several authors [5]–[7]. They use a local MPC for each entity to implement receding-horizon optimization. Other authors use a divide-and-conquer approach [8], [9]. They decompose the centralized optimization problem into many smaller optimization problems executed by local agents. Each agent can exchange information with its network neighbors. After the information is processed, agents adjust production of their DER units in a distributed manner with limited communication among the entities.

A common shortcoming of all these approaches to distributed economic dispatch is that they require prior statistical information on all DER units and loads. In addition, they cannot effectively cope with the dynamic nature of power transactions that occur under varying load and generation conditions, and at different locations.

Reinforcement learning (RL) is a powerful tool to solve complex sequential decision-making and control problems. RL can effectively learn optimal stochastic policies, even in high-dimensional or dynamic action spaces. It can reach the goal state in a few steps, with high probability, and without relying on prior information or complex stochastic modeling. These properties make RL a suitable tool to address the multistate stochastic optimization problems in modern distribution grids. As a result, RL has been widely used for energy management and demand response schemes [10].

An approach to distributed optimization in distribution systems that uses tabular Q-learning is presented in [4]. In this method, RL only finds a feasible region that contains DERs that are implicitly considered optimal. However, it does not find DERs that can contribute to the acceleration of the optimization process. In addition, tabular Q-learning does not work well with continuous observations in complex systems with many DERs. A deep RL has been adopted for real-time energy management, but only at individual home level [11]. A cooperative RL approach for distribution systems has been proposed by Liu et al. [12]. The authors suggest that each distributed controller exchanges information with its neighbours, makes action decision based on its own state and the neigbourhood states, and performs so called distributed cooperative mechanism. However, the system observability is limited to the neighbouring buses, leading to limited power transactions. In addition, this approach does nor consider the real-time impact of the line flow variations due to the power transactions.

To resolve this issue, the capability of distributed OPF algorithms has to be expanded. In addition, to deal with complex distribution circuits in stochastic environment, it is necessary to monitor network states and communicate them among the network buses. This can be accomplished through the proposed multi-agent system (MAS) architecture. This article proposes a multi-agent RL system that allows agent controllers to adapt to changes in the power distribution network as a means to maintain system security. The feasible region in a large system is obtained using Monte Carlo Tree Search (MCTS) to divide the network into multi microgrids. It uses RL to navigate from a buyer bus through the entire network (i.e. beyond the local neighbourhood). It is then followed by deep RL-based optimization procedure that finds the most suitable DER units to buy power from, while reducing the search space compared to the centralized OPF.

The uncertainty of load can substantially affect the system loss computations and the DER prices in this stochastic problem [13]. Recently, there have been several probabilistic approaches proposed to deal with this issue. Zeng et al. [14] use the regret-matching (RM) technique to analyze and correct the estimation of humans' decision-making with incomplete system information. Its stochastic optimization is solved using genetic algorithm based Monte Carlo simulation (MCS). Another possibilistic method presents a hybrid particle swarm optimization/genetic algorithm for PEVs' load modelling [15]. In this approach, uncertain factors such as home arrival time, daily travelled distance and home departure time, are based on approximating given probability distribution functions (PDFs). Uncertain wind and solar models are solved using multi-objective interval optimization [16]. This approach predicts the intervals [17] of the uncertain wind and solar power generation amounts.

However, most of these approaches rely on PDFs or MCS which average a number of simulated scenarios. For instance, MCS selects a DER unit with the highest probability in most simulated scenarios, but it might not be the right choice in some other scenarios. Thus, a few scenarios in the simulated model may impact the optimization result. The proposed approach uses deep RL that is based on advanced experiential learning. Although it is a probabilistic method, it mimics a massive number of actions to understand the system states. Unlike MCS that just averages simulated scenarios, deep RL has a powerful and robust architecture; it uses a regression process based on neural network (NN) to correlate each scenario with a best action result. Eventually, this process builds an expert system for every particular power system model. Therefore, unlike the MCS, deep RL's result is not symmetric over the load scenarios. There are almost as many unique actions strategy as there are distinctive scenarios.

In practice, operation of power systems relying on machine learning may be affected by approximation errors [18]–[20]. This may increase the cumulative operation costs of the system or even cause damage to the equipment connected to the circuit. An obvious approach to adapt deep

RL methods such as DQN to continuous domains is by simply discretizing the action space. However, this approach has a critical limitation – the curse of dimensionality: the number of actions increases exponentially with the number of degrees of freedom [21]. There are algorithms to deal with this challenge, such as deep deterministic policy gradient (DDPG) and soft actor-critic (SAC). However, manual tuning of their hyperparameters may degrade the performance. This problem has been tackled by a modified version of SAC that automates the process of selecting the optimal hyperparameters [22]. However, this algorithm is still very demanding, as it sometimes requires up to 10 million environmental steps to achieve successful training [23]. In power complex system environments, such as in multi-microgrid systems with high penetration of DERs considered here, the use of the modified SAC is impractical. It is computationally very expensive, as it requires to train every DER in each microgrid to large control steps within the DER generation capacity. Practically, the number of DERs may reach the order of hundreds in some microgrids. On the other hand, the use of exact optimization methods in complex distribution systems with stochastic DER units is often impractical due to the increase of computational burden associated with such methods. For example, the use of linear programming (LP) may not be possible due to a massive number of control variables and associated conditional statements [24]. Hence, the proposed model presents a hybrid approach that avoids the drawbacks of both constituents: machine learning errors and lack of scalability of conventional optimizers. The proposed system, called Multi Leader-Follower Actors under Centralized Critic (MLFACC), can fully capture the environment states and learn from the behavior of network participants to determine the optimal DERs before they are sent to LP for power generation optimization.

Recently, the use the alternating direction method of multipliers (ADMM) algorithm gain popularity. It breaks complex optimization problems into smaller, distributed optimization sub-problems that workable with partitioning of electric power networks. ADMM is widely used for the transmission systems, because the boundaries of the split areas are always fixed inside the main network, and their expansion in the short term is unlikely [25], [26]. In the last few years, many studies on distributed OPF algorithms have started to use the ADMM for distribution networks as well. Similar to the transmission networks, these studies assume that distribution networks are static and not affected by changing grid configurations [27]–[29].

However, modern grids usually involve a high number of DER units and load nodes that are stochastic in nature. As a result, restricting the power generation and load values in fixed zones is very challenging and it may lead to a suboptimality of DER power dispatch. The contribution of these DERs, including photovoltaic (PV) systems, electric vehicles (EV), and battery storage systems (BESS), in new power distribution systems will only increase. They induce uncertain load and generation power over the network buses, and they cannot be specified in regions with stationary boundaries inside the network. This is especially true for EVs that regularly travel between different regions/microgrids. On the other hand, suppose that there is a substantial load located very close to a boundary between two neighbouring microgrids. From the economic perspective, it may be desirable to allow this load to be supplied from both microgrids; hence, these microgrids are merged, so that all their DERs can be utilized, depending on the actual situation of the system in any given moment.

To address the issues described in the previous paragraph, the network partitioning may need to be dynamic, allowing real-time adjustments.In other words, some previously divided regions may need to be merged or reformed. Therefore, under such dynamic network partitioning, the use of ADMM technique may encounter significant challenges. Its convergence rate relies on the choice of a problem-dependent penalty factor ρ. The structure of this factor is based on a vector of variables common between the partitioning zones. The common variables are, in turn, chosen based on the power flow model of a particular network. The penalty factor also controls power flow mismatches; active power, reactive power, and the bus voltages that are used in the optimization problem constraints [30]. From a dynamic network partitioning perspective, these issues make the tuning of ρ very difficult and ADMM convergence cannot be guaranteed. All in all, the use of ADMM with a conventional partitioning method in distribution networks with high penetration and stochastic DERs that can not be restricted in one particular zone, may become impractical.

In this work, a novel, more general distributed algorithm is proposed to better accommodate the dynamic partitioning and the stochastic nature of DERs. In this algorithm, the original non-convex power flow equation for the distribution network is convexified first, then decomposed into multi-microgrid sub-problems with a dynamic partitioning ability. The proposed MCTS-RL and Dynamic Distributed Multi-Microgrid (DDMM) techniques can change the microgrid boundaries dynamically in real-time, while tracking the original network's power flow computation to guarantee its security level. Hence, these techniques can play a fundamental and crucial role in the subsequent optimization and operation of multi-microgrid systems integrated with stochastic DERs.

B. CONTRIBUTIONS

This article is primarily concerned with power distribution networks with high penetration of DER units. It highlights the necessity of building a fully distributed OPF for distribution systems that operate in stochastic environments. The major contributions of this paper are:

1) Addressing the complexity of distribution systems with high penetration of stochastic DER units through a newly proposed model called Multi Leader-Follower Actors under Centralized Critic (MLFACC). This approach facilitates cooperative interaction between all

DER units, beyond the local neighbourhood. At the same time, it maintains system security limits.

2) Resolving the suboptimality problem of distribution systems with high penetration of stochastic DER units due to the existence of substantial loads close to the boundaries between independent microgrids. The proposed combination of Monte Carlo tree search based reinforcement learning (MCTS-RL) and Dynamic Distributed Multi-Microgrid (DDMM) algorithms provides a new, flexible way to dynamically partition the network and make the system optimization and operation more efficient.

3) The proposed MLFACC algorithm accelerates the linear programming optimization method by reducing the number of arithmetic operators and their conditional statements. In effect, this simplifies the optimization problem by reducing the massive number of DERs, which may reach the order of hundreds in some real-world distribution networks.

II. POWER FLOW LINEARIZATION

In large distribution networks with high penetration of intermittent DER units (such as photovoltaic and wind generators), the power flow computational burden becomes substantial. In addition to the impact of scale, OPF needs to be checked more frequently due to the dynamic behaviour of DER units. However, the OPF in AC systems is a nonlinear, non-convex problem [31]. Therefore, finding feasible solutions for such problems is a very difficult task. A common approach is DC OPF approximation that leads to a convex optimization problem which can be solved quickly. However, its use for practical large distribution networks with a high system R/X ratio negatively affects the accuracy of OPF computations.

Yang et al. [32] illustrate the impact of several approximations used in the linearization process on branch power flows. They start from the well known polar AC power flow model

$$P_i = \sum_{j=1}^{N} G_{ij} V_i V_j \cos\theta_{ij} + \sum_{j=1}^{N} B_{ij} V_i V_j \sin\theta_{ij} \tag{1}$$

$$Q_i = -\sum_{j=1}^{N} B_{ij} V_i V_j \cos\theta_{ij} + \sum_{j=1}^{N} G_{ij} V_i V_j \sin\theta_{ij} \tag{2}$$

where N is the bus number, and G_{ij} and B_{ij} are the conductance and susceptance of the line. There are three main approximations [32] of the expression for branch power flow $G_{ij}V_i(V_i - V_j cos\theta_{ij}) \approx$

$$\text{a) } 0, \quad \text{b) } G_{ij}(V_i^2 - V_j^2), \quad \text{c) } G_{ij}(V_i - V_j).$$

Based on voltage computation results [32], the third approximation (c) provides the best accuracy. Using this simplification, the linearized models of the active and reactive power injections at bus i are [33]

$$P_i = \sum_{j=1, j\neq i}^{N} \frac{k_{ij_2}}{x_{ij}}(\delta_i - \delta_j) + \frac{k_{ij_1}}{x_{ij}}(V_i - V_j), \tag{3}$$

$$Q_i = \sum_{j=1, j\neq i}^{N} -\frac{k_{ij_1}}{x_{ij}}(\delta_i - \delta_j) + \frac{k_{ij_2}}{x_{ij}}(V_i - V_j), \tag{4}$$

where

$$k_{ij_1} = \frac{r_{ij}x_{ij}}{r_{ij}^2 + x_{ij}^2}, k_{ij_2} = \frac{x_{ij}^2}{r_{ij}^2 + x_{ij}^2}. \tag{5}$$

To solve equations (3) and (4), the node voltages have to be obtained first

$$\begin{bmatrix} P' \\ Q' \end{bmatrix} - \begin{bmatrix} B_2^c \\ -B_1^c \end{bmatrix} \delta_1 - \begin{bmatrix} B_1^c \\ -B_1^c \end{bmatrix} V_1 = \begin{bmatrix} B_2' & B_1' \\ B_1' & B_2' \end{bmatrix} \begin{bmatrix} \delta' \\ V' \end{bmatrix}, \tag{6}$$

where P', Q', δ', and V' are vectors of real power injection, reactive power injection, voltage angle, and voltage magnitude, respectively. Matrices B_1^c, B_2^c, B_1' and B_2' can be found in [33].

In a large scale power system, losses can be quite significant and their impact on the OPF and locational marginal price (LMP) cannot be ignored [33]. The flow losses for line l can be determined as follows [4], [34]

$$P_{\text{loss},l} = \frac{P_{\text{flow}}^2 + Q_{\text{flow}}^2}{V_l^2} r_l, \; Q_{\text{loss},l} = \frac{P_{\text{flow}}^2 + Q_{\text{flow}}^2}{V_l^2} x_l, \tag{7}$$

where P_{flow} and Q_{flow} are the real and reactive power flow, respectively. The loss factor is defined as a linear sensitivity of the total system losses to the real power injections at each bus with connected DER, i.e. LF $= \partial P_{\text{loss},l}/\partial P^{\text{DER}}$. Substituting (7) for $P_{\text{loss},l}$, one gets

$$\text{LF} = \left(2P_{\text{flow}}\frac{\partial P_{\text{flow}}}{\partial P^{\text{DER}}} + 2Q_{\text{flow}}\frac{\partial Q_{\text{flow}}}{\partial P^{\text{DER}}}\right)\frac{r_l}{V_l^2}. \tag{8}$$

Assuming that the reactive power is constant during DER power transactions, its derivative is zero and thus the second term can be excluded from formula (8), reducing it to

$$\text{LF} = 2P_{\text{flow}}\frac{\partial P_{\text{flow}}}{\partial P^{\text{DER}}} \cdot r_l. \tag{9}$$

III. DESCRIPTION OF THE ALGORITHM

A. MONTE CARLO TREE SEARCH BASED REINFORCEMENT LEARNING

RL and game theory can be used to develop optimization strategies for stochastic games. If considered a stochastic game, the problem of integrating intermittent, weather-dependent DERs on the grid can benefit from the developments in these areas. In conventional strategies, description of the system must be programmed in advance with sufficient prior knowledge. However, in a dynamic environment with stochastic behavior, the system itself changes over time

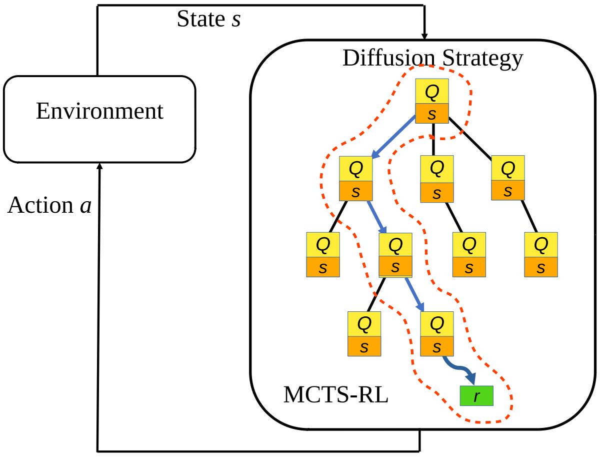


FIGURE 1. Proposed MCTS-RL algorithm.

making the optimization problem very hard to solve. In such situations, the optimization strategy can be developed by an agent through a learning process, without being explicitly programmed.

In a power system with high penetration of DERs, let S be a finite or infinite set of environment states. Each state $s \in S$ is a vector that refers to the current status of a DER unit in the search space. An agent may take an action $a \in A$ from a set of all possible actions A. The transition probability p determines the likelihood of the agent traversing from state s to s' under the joint action of all agents. In response to action a taken and state s traversed, the agent will receive an immediate local reward $r(s, a, s')$ [35]. Eventually, the learning objective of the agent is to maximize the discounted cumulative reward at each time step as follows

$$R(t) = r(t+1) + \gamma r(t+2) + \gamma^2 r(t+3) + \dots, \quad (10)$$

where $\gamma \in [0, 1]$ is a discount factor expressing the effect of the current decision on the long-term reward. A small value of γ means that rewards in the near future are more important.

Applied to power systems, feasible regions with suitable energy resources can be identified using Monte Carlo tree search-based Reinforcement Learning (MCTS-RL) [36]. This search algorithm provides the proposed approach with the ability to navigate through the power network and gradually build experience.

The regions feasible for power transactions with optimal power transfer trajectories to the DERs are determined using the diffusion strategies illustrated in Figure 1. Each bus in a power network is modeled as a node in the MCTS graph [37]. Each edge stores a set of parameters: the state-action pair (s, a) and the visit count $N(s, a)$. A learned strategy is represented by a Q-value function that maps each state-action pair to a value estimating goodness of the action in the next state s'. The Q-value function is obtained as follows

$$\begin{aligned} &Q(s_t, a_t) \leftarrow \\ &Q(s_t, a_t) + \alpha\Big[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\Big], \quad (11)\end{aligned}$$

where α is the learning rate which controls the extent of the value function update.

The next joint action is selected by the ε-greedy policy

$$a = \begin{cases} \max Q(s, a) & \text{with probability } 1 - \varepsilon \\ \text{random } a \in A & \text{with probability } \varepsilon \end{cases} \quad (12)$$

where $\varepsilon \in [0, 1]$ is the exploration rate used to balance the exploration and exploitation policies during the process of learning the Q value function. This way, the state tree is randomly built up and the experience accumulated in each state is updated by random sampling and stored in the node states by a back propagation process.

The diffusion strategy is also used to develop the partitioning method for the distribution network. It identifies buses within zones that are electrically cohesive in terms of electrical distance [27], [38]. The electrical distance theory intends to avoid paths with high impedance that result in large phase angle changes in the power flow network model. From a power transaction perspective, large phase angle changes lead to the increase of transaction leakage between buses or even between zones. In addition, MCTS-RL also considers the bus importance through their output power and demand. Since the reward function plays an important role in guiding the algorithm for the desired behavior, the reward function is designed through the system's power centroids – load or generation buses that are substantial compared to other regional buses in the network. Power centroids can be represented as [39]

$$P_c = P_j / Z_{ij}, \quad (13)$$

where P_c is the power centroid and P_j is the power generation or demand of the bus j under the MCTS-RL search space. Z_{ij} is the impedance between that bus and the root node i of the tree. This is the first stage of microgrid reformation, also called microgrid initiation.

Since bus demands change in real-time and across locations, substantial demand may occur at buses that are located close to the coupling links between the neighboring microgrids. In such cases, a microgrid that has terminal buses with heavy demand may draw power from its internal DERs that cause power losses higher than if it were connected to DERs from the neighboring microgrid. Hence, it may be better to use the DERs of both microgrids to guarantee an effective optimization result. Therefore, the second stage of the MCTS-RL algorithm is designed to check and monitor the dynamic load importance in a microgrid. This load is also called a load centroid, expressed as

$$L_c = \sum_{j \in \Omega_{MG}} P_{L,j} / P_{L,max} \quad (14)$$

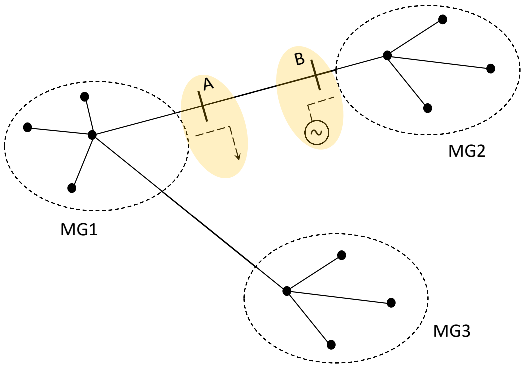


FIGURE 2. An illustrative example of microgrids' coupling-decoupling.

where L_c is the load centroid and P_L is the power demand of the inspected bus in microgrid Ω_{MG} with maximum demand $P_{L,max}$. In addition, the algorithm checks the power balance in the generated microgrid: the nominal output power of all DER units should be equal to or higher than the microgrid demand. DDMM algorithm receives all these updates and keeps tracking the newly generated microgrids to guarantee a legitimate power flow computation. Further details about the algorithm are provided in the next section.

B. DYNAMIC DISTRIBUTED MULTI-MICROGRID (DDMM)

Figure 2 shows two tie-lines that provide coupling between three neighbouring microgrids. To establish a flexible, generic dynamic decoupling for these microgrids, two conditions must be satisfied: (i) there must be a virtual decoupling method implemented, and (ii) the system must be authorized to activate and deactivate virtual decoupling for any line across the entire circuit to form or merge microgrids. To decouple a tie line within the circuit, the power injection and power flow have to be reformulated as follows:

1) Power injection decoupling: In DDMM memory, boundary bus (A) at the first microgrid is flagged as a PQ-type, whereas boundary bus (B) in the second microgrid is flagged as PV-type. Otherwise, these two microgrids are coupled. Similarly for the other neighbouring microgrids [40]–[42].
2) Power flow decoupling: equations of the linearized active (3) and reactive (4) power flows rely mainly on two independent variables: the voltage magnitude V' and voltage angle δ'. When the examined buses are located at the microgrid boundary, the voltage magnitude and voltage angle are called the boundary variables. Examination of the boundary variables is very important to prevent any violation that may jeopardize the system security, such as drawing an excessive power by one of the neighbouring microgrids from the other. Thus, the information on any changes to the boundary variables has to be provided to the DDMM algorithm.

Each microgrid agent first attempts to optimize the DER generation levels within its microgrid boundary. However, communicating through the DDMM, each microgrid also tracks its impact to the entirecircuit. This way, microgrid agents can update the estimates and, accordingly, they can change their optimization policies to maintain the required security level of the circuit. Eventually, the DDMM algorithm can be used to control the interaction between the neighbouring microgrids. The interactions of the microgrids are terminated when they reach an agreement on the amounts and prices of power supplied by their DERs. This agreement is known as consensus dynamics [43].

To enhance the learning capability in terms of DER optimization in complex power systems, the described so far; agent-based algorithm can be expanded to multiagent case through the proposed MLFACC method. A theoretical framework of this method is introduced in the next section.

C. MULTI LEADER-FOLLOWER ACTORS UNDER CENTRALIZED CRITIC

The proposed MLFACC algorithm relies on deep reinforcement learning using the Advantage Actor Critic (A2C) algorithm [44]–[46]. A2C is the best fit for the proposed distributed optimization algorithm. Three actor networks train decentralized policies in a multi-agent framework, and share information using a centralized critic network. The main idea of using a critic network is to learn a centralized policy with an attention mechanism. In complex multi-agent environments, the attention mechanism has shown effective and scalable learning [47]. The intuition behind this idea is that the centralized critic can dynamically evaluate each agent's action; eventually, it sends attention to the agents to adjust their actions according to the environment need.

Another crucial approach to obtain the optimal variables of interest to accelerate the LP method is the leader-follower policy. The idea of the leader-follower game policy is inspired by Stackelberg game model [48]. In order to take an optimal action, it is necessary that a leader fully understand the environment and not only learns from its own actions but also the follower's actions. Typically, the leader acts first, then announces its action. At this point, the game rule allows the followers to make their decisions. In the proposed method, the roles of the players in the game change: if the number of agents is more than two, every follower agent can be a leader to the next agent. However, the first agent is always a leader, and the last agent is always a follower. Also, it is worth noting that the follower's action is estimated as a function of the leader's actions since the goal's reservation of the previous leader is already made. Thus, in this game, the leader uses a competitive policy, while the follower is expected to use a cooperative policy.

The main question that arises in this algorithm is how agents learn from each other the optimal policies and get higher rewards. The simplest form of policy gradient method is REINFORCE which represents gradient as [35], [49]

$$g = \mathbb{E}_{s_{0:\infty}, a_{0:\infty}}[\sum_{t=0}^{T} R_t \nabla_\theta \log \pi_\theta(a_t, s_t))]. \tag{15}$$

Policy π_θ is trained by following the gradient that relies on a critic network, which estimates the value function. In particular, R_t is replaced by any expression equivalent to $Q(s_t, a_t) - b(s_t)$, where $b(s_t)$ is a baseline designed to reduce the variance. Common options are to substitute $v(s_t)$ for $b(s_t)$, and to replace R_t by the temporal difference (TD) error $r_{t+1} + \gamma v(s_{t+1}) - \hat{v}(s_t)$ [49]. Term $\hat{v}(s_t)$ is the predicted or approximated value of value network. It is computed by a multi-layer NN, with a vector of connection weights in all layers θ. The target value $r_{t+1} + \gamma v(s_{t+1})$ is obtained from the immediate reward r_{t+1} and the discounted estimated value of next state $\gamma v(s_{t+1})$. To estimate the error between the approximated value and the target value, stochastic gradient descent (SGD) is used. The approximate value function $v(s, \theta)$ is a differentiable function of θ for all $s \in S$ [35], [50]. If an agent performs an action, it is based on the states mapped through the critic network. These states are always changing by the actions of the agent itself as well as the other agents. In other words, all agents should take optimal policies by their action probabilities as in $(\nabla \log \pi_\theta(a_t, s_t))$ in order to increase the return in the critic network in $(r_{t+1} + \gamma v(s_{t+1}, w_{t+1}))$ of the same equation. To represent this, (16) can be formulated as the following by basing on the previous equations:

$$\nabla_\theta J_{\pi\theta}^{\mathcal{E}} \leftarrow \nabla \log \pi_\theta(a_t^{\mathcal{E}}, s_t^{\mathcal{E}})(r_{t+1} + \gamma v(s_{t+1}, w_{t+1}) - \hat{v}(s_t, w_t)), \tag{16}$$

where index $\mathcal{E}$ refers to a particular agent under policy estimation. To reduce the variance of value functions, the advantage function is estimated by the TD-error:

$$A = r_{t+1} + \gamma v(s_{t+1}) - \hat{v}(s_t). \tag{17}$$

Since agents seek their own, unique goals, each agent has its own loss gradient $(\nabla_\theta J_{\pi\theta}^{\mathcal{E}})$ that is sent to the critic network to allow estimation of the policy probabilities advantage $(\hat{A}_t)$. Substituting (17) in (16), the final expression can be represented as:

$$\nabla_\theta J_{\pi\theta}^{\mathcal{E}} \leftarrow \nabla_\theta \log \pi_\theta(a_t^{\mathcal{E}}, s_t^{\mathcal{E}}) \hat{A}_t. \tag{18}$$

This way, the central critic can teach the agents based on the experience of the other agents and the state updates of the system. To update the parameters of the policy network θ a gradient descent of the SGD rule is used:

$$\theta_{t+1} \leftarrow \theta_t + \alpha \nabla_\theta J_{\pi\theta}^{\mathcal{E}}, \tag{19}$$

where α is the learning rate for the actor network, and the gradient $\nabla_\theta J_{\pi\theta}^{\mathcal{E}}$ is the gradient calculated by (18). Similarly for updating the parameters of the critic network:

$$w_{t+1} \leftarrow w_t + \beta \delta \hat{v}_{(s,w)}. \tag{20}$$

To prevent the follower agents from seeking the same leader's policy trajectories, a tracing constraint $(\nabla_\theta \log \pi_\theta(a_t^{\mathcal{L}}, s_t^{\mathcal{L}}))$ is added to (18), as

$$\nabla_\theta J_{\pi\theta}^{\mathcal{E}} \leftarrow \nabla_\theta \log \pi_\theta(a_t^{\mathcal{E}}, s_t^{\mathcal{E}}) \hat{A}_t - \mu\big[\max\big(\nabla_\theta \log \pi_\theta(a_t^{\mathcal{L}}, s_t^{\mathcal{L}}), \Psi\big)\big], \tag{21}$$

where $\mathcal{L}$ is the index of all agent policies in a leader position, and μ is a lagrangian multiplayer. This constraint forces the follower agent that intends to choose the best goal to be right inferior to the leader's goal. However, this constraint may result in inefficient policies by the follower agents and slow learning. Since the leader plans its strategy to propel the followers to take actions in its favour, it may pick a trivial trajectory; consequently, the followers are constrained to choose other trajectories with even less importance. Hence, to relax this constraint at the beginning of the learning process, a relaxation factor (Ψ) is used. This factor is also an action probability that allows the followers to break the tracing constrained to a particular limit; once the leader finds a proper trajectory that leads to obtain a better DER price, the relaxation factor vanishes. At this point, just the first term of the maximization operator is valid. This mechanism enables the proposed algorithm to identify the best group of DERs in a descending order, and without overlap.

The reward function originates from the DER units that have the minimal active power flow losses. Based on their locations, their engagement may reduce the active power flow losses in a distribution network. Therefore, the reward function is formulated as

$$R = \min \sum_{i \in \Omega_k} f_i(C_i^{\text{losses}}), \tag{22}$$

where Ω_k refers to the feasible region that is generated by MCTS, and index i the index of DER units. The agent states include all conditions required to make an appropriate decision, including all relevant power system constraints. More details about the system constraints are provided in the next section which illustrates the optimization of the DER engagement using RL-based Linear Programming. The pseudocode is presented in Algorithm 1.

To explain the operation of the MLFACC algorithm, assume that penetration of DERs, and especially EVs, is 30%. To model this demand, it can be considered a load centroid uniquely randomly distributed across all scenarios of a stochastic game. Further, it is assumed that each load centroid correspond to 150% of a particular bus-load in a microgrid. Through its learning policy, MLFACC attempts to identify DERs with the lowest power generation prices. The price differentiation of the DER power generation units is inversely proportional to their power losses. However, uncertainty of load locations can substantially affect the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3075247, IEEE Access

Algorithm 1: MLFACC algorithm

```
Initialize actors' weights; $(\theta_1, ..., \theta_N), \theta \in \mathbb{R}^n$
Initialize critic weight; $(w)$, $w \in \mathbb{R}^m$
Initialize step size parameters: $\alpha > 0,\ \beta > 0$.
for $t$ = 1 to max episode length do
    done = False
    while not done do
        for agents $i$ = 1 to $N$ do
            Choose action $a_{t+1}$ based on probability:
            log $\pi_\theta(a_t, s_t)$
            Receive observation $(s_{t+1}, s_t, r_{t+1}, done)$
        end for
        Compute the TD error:;
        $\delta \leftarrow (r_{t+1} + \gamma\hat{v}(s_{t+1}, w_{t+1}) - \hat{v}(s_t, w_t))$;
        Compute the loss gradient for each agent by
          eq. (21);
        Update policy parameters for the actor
          networks:
        $\theta_{t+1} \leftarrow \theta_t + \alpha\nabla_\theta J^{\mathcal{E}}_{\pi\theta}$
        Update policy parameters for the critic
          network:
        $w_{t+1} \leftarrow w_t + \beta\delta\hat{v}_{(s,w)}$
    end while
end for
```

computation of system losses and, consequently, the prices of DER units.

The proposed MLFACC algorithm plays this stochastic game, represented by an interactive environment between the random load centroids and stochastic behavior of the DERs. Specifically, MLFACC uses the power flow calculated from the network model to find the optimal DER candidates with the lowest power losses that do not violate any grid constraints (such as voltage limits and line congestion). To determine the best candidates, the DER selection trials are simulated by the actor-network actions of the algorithm. Although it is a probabilistic method, the selected DER are not symmetrically distributed over the load scenarios. Unlike Monte-Carlo and other probabilistic methods that are based on averaging the data, the DNN of the actor-critic network in the MLFACC architecture is based on a non-linear regression analysis. In other words, the DNN performs a correlation process between the outputs (labels) and the input data (load scenarios). Hence, virtually every load scenario receives a unique result.

To ensure a successful correlation process when analyzing system states, the sample efficiency has to be considered through the actor-networks of the MLFACC algorithm. This can be achieved using the advantage function (18) which also contributes to reducing the variance. Learning is initiated using an exploration policy that uses random actions to perform preliminary examination of the system state. In addition, the actor-networks under the policy gradient method also perform a number of deterministic exploitation actions. In this step, each new action works along an existing action with the aim to perform behaviors that yield better results. At every epoch, the actor-networks collect news experience that is sent to the critic network. The critic-network continuously updates its weights, attempting to find the correlation between the input data and the targets. This process continues until the network converges to the final result.

The strength of this approach becomes clear when applied to real-world distribution networks with hundreds of DERs. In such cases, optimization of DERs dispatch using conventional methods becomes very challenging. Using the MLFACC approach that employs deep RL, the stochastic game results in selecting a small number of DERs as the best candidates for the subsequent LP optimization. Nevertheless, if a substantial load centroid is located close to the coupling line connecting two neighbouring microgrids, the stochastic problem turns into a deterministic one, and the use of the MLFACC is no longer needed to find the best DERs for these bus locations.

D. DEEP RL-BASED LINEAR PROGRAMMING

A power distribution network can be modeled by a directed tree graph $T(\Omega_{MG}, \Omega_L) \subseteq (N, L)$. The nodes of the graph Ω_{MG}, a subset of all network buses N, are linked by a set of distribution lines Ω_L, a subset of all network lines L. Node 0 is the starting point of the tree search, referred to as the root node j. In general, the root node can represent either a buyer or a seller of energy. Under the scenario considered in this study, the root node is specified as a buyer looking for the best seller(s). The remaining nodes are referred to as branch nodes. Each pair of adjacent branch nodes is connected by a branch line $l \in \Omega_L$. All nodes (except the leaf nodes) in this tree are parent nodes since they have a set of child nodes C_i linked by the branch lines. In addition, the child nodes may have connected DER units. i is the index of all buses that link to load bus j. Each line in L has an impedance $z_i = r_i + x_i$. Power injection from node i to node j is calculated using equation (3).

It is worth noting that there are many factors that significantly impact deep RL accuracy. These factors include the number of DERs, random variables such as the random load centroid, the non-linearity of the system, the number of system constraints, the resolution of time-series that may involve massive generation and/or demand variations during the day, and finally the circuit size. These factors leverage the relatively more complex relationships in the data of the system states and DERs' generation amounts for deep RL training. To reduce this complexity, the optimization of the DER power generation amount within each microgrid is eliminated from the deep RL decision task and, instead, it is determined deterministically. Therefore, to minimize the cost of DER generation dispatch under system constraints, we propose a new distributed OPF algorithm based on deep learning called Deep Reinforcement Learning-based

Linear Programming (DRL-LP). In this model, an optimal power generation that is determined by LP is accelerated by selecting the optimal DERs through MLFACC, determined within each microgrid by MCTS-RL and DDMM. Chazelle and Matousek [24] have analysed and estimated the computational complexity that describes the amount of time it takes to run LP by counting the number of input variables x and $g(x)$ constraints as follows

$$O(x)^{7x}(\log x)^{x} g(x), \tag{23}$$

where $O(.)$ denotes the time complexity. Thus, the behavior of the LP complexity can be reduced by reducing the size of the input.

Based on the linearized power flow model described in section II, the DRL-LP problem can be formulated as follows

$$\min \sum_{x \in \text{MLFACC}_{MG}} f_x(C_x^P), \tag{24}$$

$$g(x) = \begin{cases} P_i - d_j = \sum\limits_{i \in \Omega_{MG}} P_{ij}^{\text{flow}} + P_{ij}^{\text{loss}}, & (25) \\ \theta_{\text{ref}} = 0, & (26) \end{cases}$$

$$h(x) = \begin{cases} \underline{P}_{\text{DER}} \leqslant P_{\text{DER}} \leqslant \overline{P}_{\text{DER}} \forall \text{DER} \in \Omega_{MG}, & (27) \\ \underline{P^{\text{flow}}}_{ij} \leqslant P_{ij}^{\text{flow}} \leqslant \overline{P_{ij}^{\text{flow}}} \ \forall l_{ij} \in \Omega_{MG}, & (28) \\ \underline{P^{\text{loss}}}_{ij} \leqslant P_{ij}^{\text{loss}} \leqslant \overline{P_{ij}^{\text{loss}}} \ \forall l_{ij} \in \Omega_{MG}, & (29) \\ \underline{V}_i \leqslant V_i \leqslant \overline{V}_i \quad \forall x \in \Omega_{MG}, & (30) \\ \zeta_{ij} \leqslant \phi \qquad \forall l_{ij} \in \Omega_{MG}, & (31) \end{cases}$$

In this optimization problem, C_i^P is the optimal DER that is determined by MLFACC. It belongs to a node in the tree graph Ω_{MG} delineated by MCTS as a feasible subset of the original network. The objective function aims to minimize the generation cost at node i, and implicitly minimizes the losses P_{ij}^{loss} of the line connecting nodes i and j. Functions $g(x)$ and $h(x)$ express, respectively, the equality and inequality constraints. The nodal balance power flow is restricted by constraint (25), where d_j is the power demand, while equation (26) holds the reference bus voltage angle at zero. Inequalities (27)–(29) express the upper and lower bounds of the power output of DER units, the power flows in the branches, and the power losses in the branches, respectively. The coupling constraint between the microgrid Ω_{MG} and its neighbors is denoted ϕ. Based on the concept of electric distance, ϕ represents the threshold value of ζ of the line impedance between the load bus j and the cross-border buses separating microgrid from its neighbors. It can be considered a means to specify the borders of a feasible space of the DOPF problem among multiple regions. However, the network constraints must still be observed and communicated among microgrids. DDMM can efficiently manage the information for multi-microgrid systems. From the implementation perspective, all information is sent to the DDMM during the distributed optimizer instantiations and load bus solutions of the OPF subproblems. The DDMM reconciles system state information for multiple microgrids.

IV. REAL-TIME BALANCING ELECTRICITY MARKET

To illustrate application of the proposed DOPF method using DRL-LP, we construct a distribution electricity market framework to facilitate the effective integration of DERs into the electricity system. A central role in this framework is assumed by the distribution system operator (DSO) who facilitates DERs integration and delivers location services. It also provides real-time power balancing through dispatch of stochastic DERs and bidding of flexible loads.

The algorithm for balancing the electricity market is executed every minute to accommodate (near) real-time power imbalances. Distribution locational marginal price (DLMP) differs from location to location due to the limits of the node voltage, line capacity, and network losses. This facilitates the mitigation of over/under voltage and line congestion, and compensation of location-dependent network losses.

The main goal of the DSO is to maximize its economic benefits while providing the amount of power required by the balancing electricity market. The individual entities of the distribution system respond to specific price signals derived from the following DLMP equation

$$\begin{aligned} \text{DLMP}_i = \lambda_0^p + \lambda_0^p \cdot \sum_{i=1}^{N} \frac{\partial P_i^{\text{flow}}}{\partial P_i^{\text{DER}}} + \lambda_0^p \cdot \sum_{i=1}^{N} \frac{\partial P_i^{\text{loss}}}{\partial P_i^{\text{DER}}} + \\ + \lambda_0^p \cdot \sum_{i=1}^{N} \frac{\partial V_i}{\partial P_i^{\text{DER}}} \end{aligned} \tag{32}$$

where λ_0^P is the active power exchange or the reference price. This is a known parameter that can be adjusted by the DSO. The three sum terms $\sum_{i=1}^{N} \partial P_i^{\text{flow}} / \partial P_i^{\text{DER}}$, $\sum_{i=1}^{N} \partial P_i^{\text{loss}} / \partial P_i^{\text{DER}}$, and $\sum_{i=1}^{N} \partial V_i / \partial P_i^{\text{DER}}$ are the total line flow factor, total system losses factor, and voltage deviation factor, respectively. All three factors are calculated with respect to DER power injection, P_i^{DER}, from the constraint equations (28), (29), and (30). DLMP works as a price coordinator to ensure that any power imbalance in the system can be fully offset and objectives of all participating entities can be optimized simultaneously. This coordinated operation model is designed to include all required objective functions and system constraints.

V. RESULTS AND DISCUSSION

To demonstrate the proposed DRL-LP algorithm, the modified IEEE 123-bus test system [51] is considered as a case study. To examine the leverage provided by the RL agents in the distributed optimization subproblems, a search tree is progressively built using MCTS-RL. This tree is a randomly biased sequence of actions applied by an RL agent to a given series of states until a predefined coupling constraint ϕ is reached. This way, the feasible region suitable for power transactions is obtained based on the concept of electric distance. The MCTS-RL process instantiates six microgrids as shown in Figure. 3. Suppose that the stochastic load centroids are randomly distributed over the system buses with a penetration of 30% of the entire microgrid buses, and the value of each load centroid is 150% of a particular bus load in a microgrid. In this simulation, it is considered that all buses in the microgrid have DERs with a limit of 100 kW each. The algorithm determines the optimal DER candidates that have the lowest power losses and maintain the grid limits, voltage limit and line congestion, in a particular microgrid. Consequently, the number of variables and their conditional statements is reduced to accelerate the LP optimization process.

To illustrate the learning simulation of obtaining a proper number of optimal DERs by the MLFACC algorithm, first microgrid is chosen, which is labeled as MG1 in Figure 3. For simplicity, three agents are used in this illustration to obtain the three best DERs in each microgrid. Figure 4, shows that most agents converged after about 5900 episodes. The following subsections extensively analyze the optimization process in three cases: (A) normal system operation, (B) a system with dynamic microgirds, and (C) a system with line congestion.

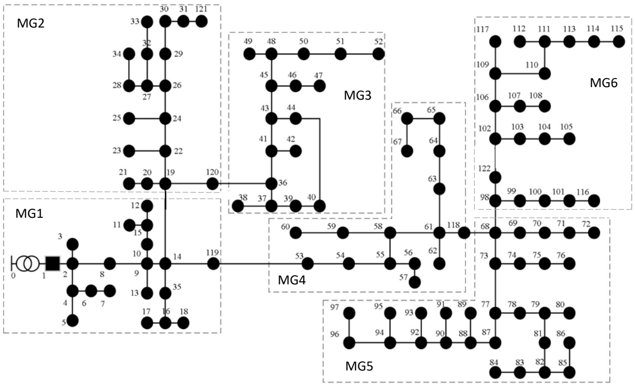


FIGURE 3. The modified IEEE 123-bus distribution with 6-microgrids

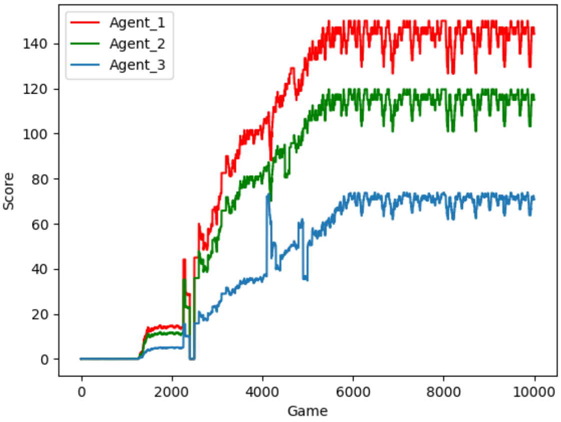


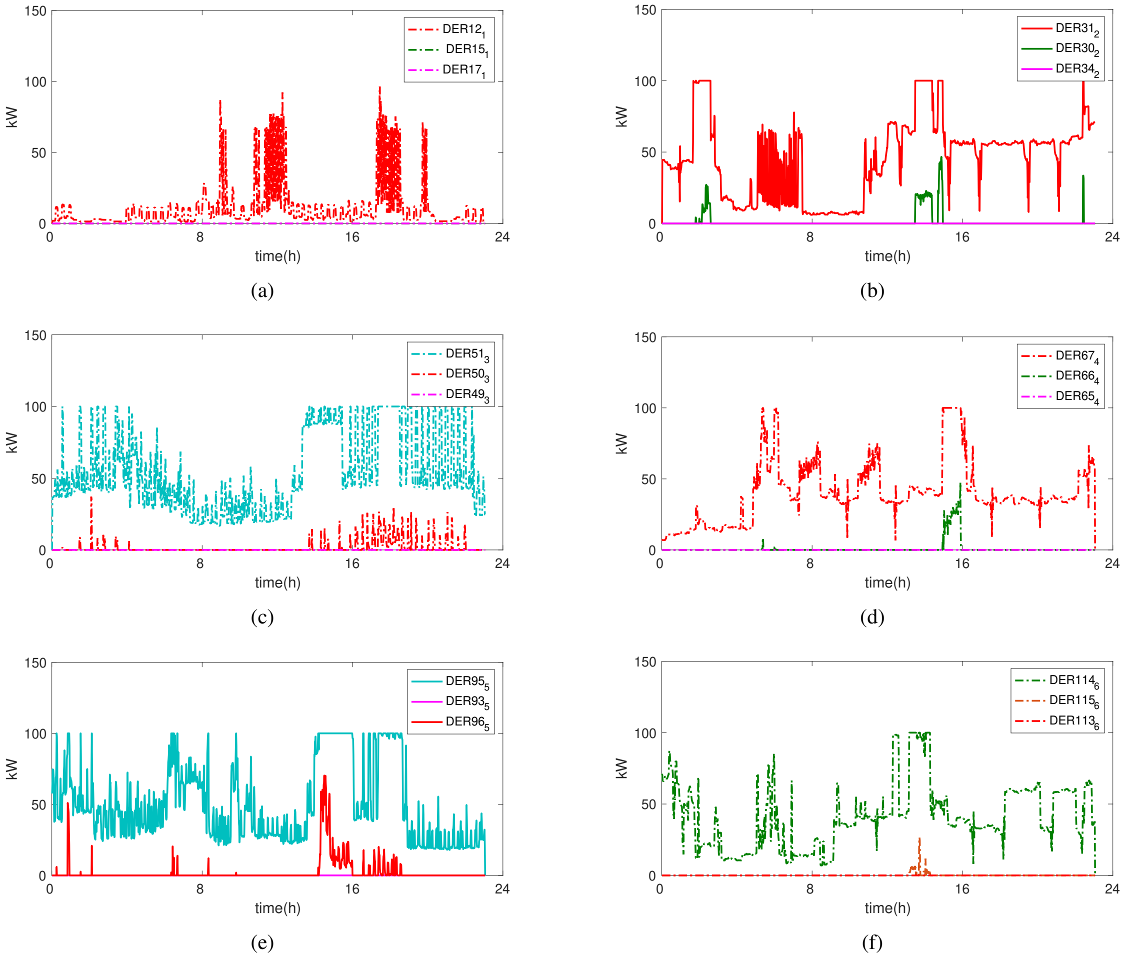
FIGURE 4. The learning simulation of the MLFACC for region A buses.

A. NORMAL SYSTEM OPERATION

The MLFACC algorithm has already been trained under the random distribution of load centroids. To test the optimization process through the algorithm, and for simplicity, a single dynamic real-time load centroid is chosen in each microgrid of the modified 123-bus system, at the following buses: 9, 23, 43, 58, 76, and 108, for the microgrids, 1, 2, 3, 4, 5, and 6, respectively. The load data has been extracted from a residential community in Edmonton, Alberta, Canada, and scaled to the transformer level. This load is considered an extra load to be balanced by the generation of the main feeder of the circuit and the DER units in each microgrid. Loads of the remaining buses of the circuit are based on the original static load data. Each microgrid operates independently and is responsible for its DERs when no merging process is exerted by the DDMM algorithm. Real-time optimization of the DERs' generations are shown in the Figure 5. The optimization is performed on a 1-minute basis, for instance: the MLFACC algorithm obtains the best DERs of the first microgrid on buses 12, 15, and 17, as shown in Figure 5.a. It can be seen that, when DER12 (at bus 12) reaches its limit (100kW), the algorithm switches to DER15. In notation DER$\#_i$, index i refers to the microgrid number. The optimization process relies mainly on the variation of DER benefits stemming from the reduction of the active power flow losses within the circuit. Based on their location, the power the DERs generate usually flows in the direction opposite to the main power stream of the feeder. This causes a reduction of the total losses of the feeder power flow. The variation of losses also causes considerable differences of DLMP, especially when the DER generation levels approach their maximum energy export capacity. More details about the DLMP pricing are provided in the third case (system with line congestion), to include the microgrids' merging and line congestion impact on the system pricing. The proposed algorithm (including all its components) executes in approximately 5 seconds to determine 1 minute of real-time optimized power generation. However, the deep RL-LP algorithm uses the experience of MLFACC obtained through training involving 10,000 games in about 84 hours required for one-time of training.

B. SYSTEM WITH DYNAMIC MICROGIRDS

In the previous case, the MLFACC algorithm was provided plenty of training time and samples to play the stochastic game and pick the most efficient and secure DER units for each microgrid. However, when the load centroids are close to the coupling lines between neighbouring microgrids,



(a)

(b)

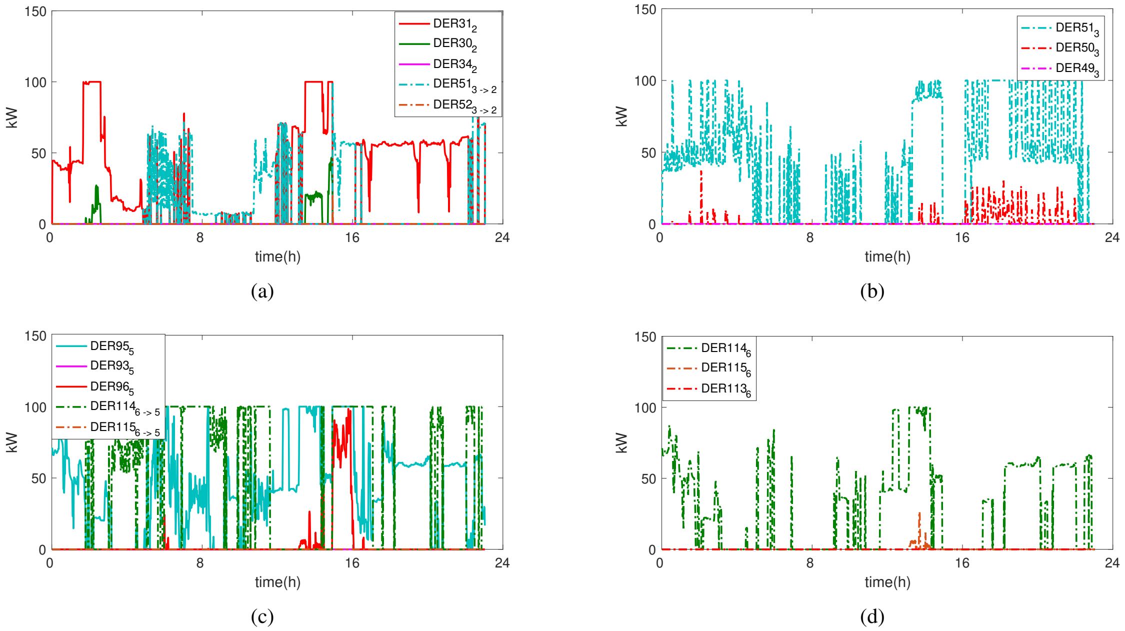
(c)

(d)

(e)

(f)

FIGURE 5. The real-time optimization of the DER generations in the microgrids (1-6), referred to in the figures (a-f), respectively.



(a)

(b)

(c)

(d)

FIGURE 6. The real-time optimization of the DER generations in the dynamic microgrids 2,3,5, and 6, referred to in the figures (a-c), respectively.

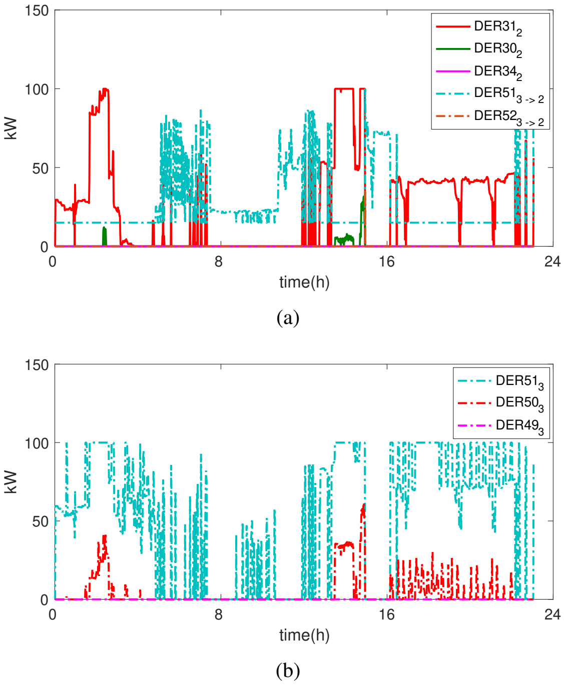
these microgrids are merged. Due to the fact that most partitioning algorithms result in only a few coupling lines between the newly generated microgrids, the optimization problem can be turned into a deterministic problem. Hence, it is easy to determine the best DERs over the microgrids surrounding the coupling lines.

In addition to the information regarding the load centroids provided in the previous case, it is assumed that there is another load centroid at bus 19 of microgrid 2. This load centroid is labeled as ($L_{c_{2-3}}$) as it is located at the coupling line between microgrids 2 and 3. Since $L_{c_{2-3}}$ is real-time load, if it reaches a high load value, it is identified as a load centroid and the two microgrids are merged. Otherwise, they are split, as this load is considered a normal load affiliated with microgrid 2. For simplicity, if the value $L_{c_{2-3}} \geq 1$, it is considered a load centroid. Figures 6.a and 6.b show the merge/split process in real-time for the microgrids 2 and 3, respectively. When they work independently, the best DERs are 31, 30, and 34 for microgrid 2, and 51, 50, and 49 for microgrid 3. When they are merged into a single microgrid, the new best DERs are 51, 52, and 49 (calculated deterministically). The power optimization of the merging process is shown in the second microgrid result, at the following time periods: 5.00–13.20, and 22.30–23.30. On the other hand, for the same time periods, microgrid 3 produces no output power to represent the the fact that the microgrid is merged with microgrid 2. In addition, the original values of DER calculated by the MLFACC of each microgrid are still considered in the optimization problem due to the presence of the original load centroids at buses 9 and 23. However, some DERs (such as 49 and 51) are common for both cases (merge/split). Similar situation is observed when considering another load centroid Lc_{4-5-6} at bus 68, which is a terminal of two coupling lines between the microgrids 4, 5 and 6. Note that, when the DERs are selected deterministically during the merging process, they are just from microgrids 5 and 6: 114, 95, and 96. Moreover, the load centroid Lc_{4-5-6} is located in microgrid 5. Since there is no participation from microgrid 4, only microgrids 5 and 6 are considered in the merging process. The optimization result of the microgrids 5 and 6 are shown in Figures 6.c and 6.d, respectively. In the same figures, when the microgrids split and work independently, the selected DERs are 95, 93, and 96 for microgrid 5, and 114, 115, and 113 for microgrid 6.

C. SYSTEM WITH LINE CONGESTION

Typically, when demand is concentrated on a few DERs, the corresponding segments of distribution lines can become overloaded. To mitigate the occurrence of overloaded lines due to DER generation, agents in each microgrid have to track their impact on system security limits. This way, the MLFACC algorithm is capable of maintaining microgrid security limits under a stochastic load environment.

The DDMM algorithm also helps preventing any limit violations across the coupling lines and the network in general by sharing this information among microgrids. Therefore, the risk of congestion that would threaten the coupling lines is very low. Thus, it is assumed that the flow limit of a selected coupling line is reduced in comparison to its original value. The line connecting buses 120–36 and the coupling between microgrids 2 and 3 are chosen for the contingency study so that the microgrid merging process in demonstrated as well. The flow limit of this coupling line, $C_{l_{2-3}}$, is reduced from 755 kW to 740 kW.



(a)

(b)

FIGURE 7. The real-time optimization of the DER generations in the dynamic microgrids 2 and 3 under line congestion, referred to in the figures a and b, respectively.

When the load of microgrid 3 increases such that the flow on line l_{2-3} exceeds 740 kW, the flow congestion occurs at this line. Typically, the amount of generation of the DERs that compromise network security is reduced. Instead, another DER that does not influence the system security, while offering an acceptable price, is called. In such case, DER unit 51 in microgrid 3 keeps its power generation even when the load centroid reduces to zero as shown in Figure 7a. This DER unit attempts to compensate for the resultant power reduction in microgrid 3 due to the new flow limit.

The primary goal of this step is to change the system flow to prevent congestion, while providing energy to the load at an acceptable price. However, this leads to a step change of price (Figure 8). The price change to avoid 1 kW of line congestion is called congestion DLMP. Similarly, the change to avoid line losses is called losses DLMP. The daily values of congestion and losses DLMP in microgrid 2 are 7.58 ¢ and -125.62 ¢, respectively. In microgrid 3, these values are 7.94 ¢ and 0.12 ¢, respectively. The negative sign of the second microgrid's losses indicates a reduction of losses due to DER generation. Conversely, the positive sign of losses DLMP for microgrid 3 indicates a decline in counterbalance of power flow losses' in the direction between the main feeder and the DERs; this due to the reduction in the main feeder power generation. The DLMP values for these two microgrids are also shown in Figure 8. Finally, the total cost of generation in these microgrids for this case is shown in Figure 9.

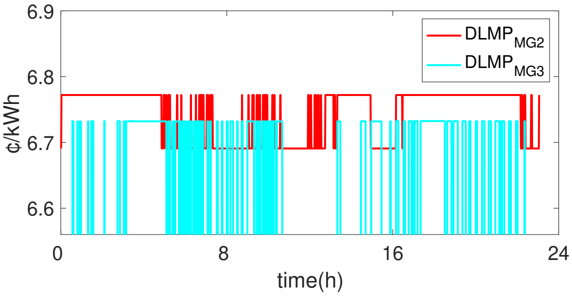


FIGURE 8. The total generation costs for both regions A and B.

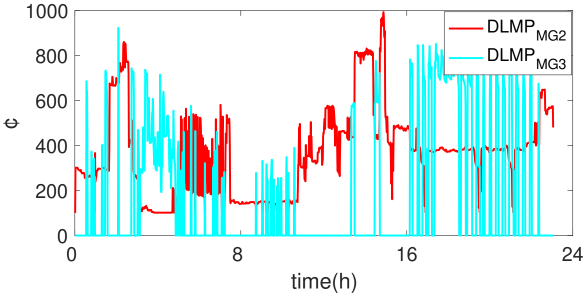


FIGURE 9. The total generation costs for both regions A and B.

VI. CONCLUSIONS AND FUTURE WORK

This article introduces a novel approach in distributed OPF for distribution systems with high penetration of DERs. Using modern methods of artificial intelligence, the proposed approach facilitates OPF calculation while reducing its computational burden. The proposed method is based on an effective combination of Monte Carlo tree search-based reinforcement learning (MCTS-RL) and the dynamic distributed multi-microgrid (DDMM) algorithm. Through the dynamic network partitioning and navigation steps of a diffusion strategy, they generate adaptable microgrid configurations with a set of optimal paths to the most suitable generation and load nodes.

The proposed deep learning-based actor-critic approach (MLFACC) mitigates the challenges associated with the stochasticity of DERs while addressing the problem of dimensionality faced by conventional optimization techniques. Only the selected DERs are then considered by the optimizer applied to the linearized problem, thus guaranteeing convergence. The multiagent nature of the proposed approach allows a direct application of DOPF in systems with multiple interacting entities.

The presented simulation results clearly demonstrate the effectiveness of the proposed method to solve the distributed economic dispatch problem while maintaining the system security limits.

The IEEE 123-bus system considered in this study experiences only minor voltage control issues. The losses are also relatively low as they are proportional to the network size. Nevertheless, the obtained results confirm that the proposed model can successfully consider dynamic microgrid configurations and provide an effective power market solution without jeopardizing system security. In addition, the proposed methodology is suitable for large and complex networks that can accommodate various DER types such as PV systems, EVs, and BESS that induce uncertain load and generation patterns [52]–[54] . A good example to illustrate this approach is an EV system model that has high complexity due to stochastic transportation patterns. Therefore, in future work, the authors will consider the EV system model in more detail, to show the potential applicability of the proposed approach to distributed optimization of such complex systems.

REFERENCES

[1] P. D. Achlerkar, S. Kewat, B. K. Panigrahi, and B. Singh, "Distributed control framework for autonomous seamless operation of electronically interfaced distributed generators in ac microgrid," in *2018 8th IEEE India International Conference on Power Electronics (IICPE)*, Dec 2018, pp. 1–6.

[2] S. K. Singh and N. Bansal, "Output impedance as figure of merit to predict transient performance for embedded linear voltage regulators," in *2014 27th International Conference on VLSI Design and 2014 13th International Conference on Embedded Systems*, Jan 2014, pp. 516–521.

[3] M. Střelec, K. Macek, and A. Abate, "Modeling and simulation of a microgrid as a stochastic hybrid system," in *2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, Oct 2012, pp. 1–9.

[4] M. Al-Saffar and P. Musilek, "Distributed optimal power flow for electric power systems with high penetration of distributed energy resources," in *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, 2019, pp. 1–5.

[5] A. Saad, T. Youssef, A. T. Elsayed, A. Amin, O. H. Abdalla, and O. Mohammed, "Data-centric hierarchical distributed model predictive control for smart grid energy management," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4086–4098, July 2019.

[6] E. R. Stephens, D. B. Smith, and A. Mahanti, "Game theoretic model predictive control for distributed energy demand-side management," *IEEE Transactions on Smart Grid*, vol. 6, no. 3, pp. 1394–1402, May 2015.

[7] Y. Zheng, S. Li, and R. Tan, "Distributed model predictive control for on-connected microgrid power management," *IEEE Transactions on Control Systems Technology*, vol. 26, no. 3, pp. 1028–1039, May 2018.

[8] Q. Li, F. Chen, M. Chen, J. M. Guerrero, and D. Abbott, "Agent-based decentralized control method for islanded microgrids," *IEEE Transactions on Smart Grid*, vol. 7, no. 2, pp. 637–649, March 2016.

[9] F. Guo, C. Wen, J. Mao, and Y. Song, "Distributed economic dispatch for smart grids with random wind power," *IEEE Transactions on Smart Grid*, vol. 7, no. 3, pp. 1572–1583, May 2016.

[10] F. Alfaverh, M. Denaï, and Y. Sun, "Demand response strategy based on reinforcement learning and fuzzy reasoning for home energy management," *IEEE Access*, vol. 8, pp. 39 310–39 321, 2020.

[11] M. Khan, J. Seo, and D. Kim, "Real-time scheduling of operational time for smart home appliances based on reinforcement learning," *IEEE Access*, vol. 8, pp. 116 520–116 534, 2020.

[12] W. Liu, P. Zhuang, H. Liang, J. Peng, and Z. Huang, "Distributed economic dispatch in microgrids based on cooperative reinforcement learning," *IEEE*

Transactions on Neural Networks and Learning Systems, vol. 29, no. 6, pp. 2192–2203, June 2018.

[13] T. Huiling, W. Jiekang, W. Fan, C. Lingmin, L. Zhijun, and Y. Haoran, "An optimization framework for collaborative control of power loss and voltage in distribution systems with dgs and evs using stochastic fuzzy chance constrained programming," *IEEE Access*, vol. 8, pp. 49 013–49 027, 2020.

[14] B. Zeng, J. Feng, N. Liu, and Y. Liu, "Co-optimized parking lot placement and incentive design for promoting pev integration considering decision-dependent uncertainties," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 1863–1872, 2021.

[15] A. Ahmadian, M. Sedghi, A. Elkamel, M. Aliakbar-Golkar, and M. Fowler, "Optimal wdg planning in active distribution networks based on possibilistic–probabilistic pevs load modelling," *IET Generation, Transmission & Distribution*, vol. 11, pp. 865–875(10), March 2017. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/iet-gtd.2016.0778

[16] F. Wei, Q. Wu, Z. Jing, J. Chen, and X. Zhou, "Optimal unit sizing for small-scale integrated energy systems using multi-objective interval optimization and evidential reasoning approach," *Energy*, vol. 111, pp. 933 – 946, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360544216306569

[17] A. Zarnani, P. Musilek, and J. Heckenbergerova, "Clustering numerical weather forecasts to obtain statistical prediction intervals," *Meteorological Applications*, vol. 21, no. 3, pp. 605–618, 2014.

[18] J. Guo and C. . Yang, "Impact of prediction errors on high throughput predictive resource allocation," *IEEE Transactions on Vehicular Technology*, pp. 9984 – 9999, 2020.

[19] R. Medar, V. S. Rajpurohit, and B. Rashmi, "Impact of training and testing data splits on accuracy of time series forecasting in machine learning," in *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 2017, pp. 1–6.

[20] N. Kulathunga, N. R. Ranasinghe, D. Vrinceanu, Z. Kinsman, L. Huang, and Y. Wang, "Effects of the nonlinearity in activation functions on the performance of deep learning models," in *2020 arXiv preprint arXiv*, 2020.

[21] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *2015 arXiv preprint arXiv*, 2015.

[22] A. Lazaridis, A. Fachantidis, and I. Vlahavas, "Deep reinforcement learning: A state-of-the-art walkthrough," *Journal of Artificial Intelligence Research*, vol. 69, pp. 1421–1471, 2020.

[23] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft actor-critic algorithms and applications," in *2018 arXiv preprint arXiv*, 2018.

[24] B. Chazelle and J. Matoušek, "On linear-time deterministic algorithms for optimization problems in fixed dimension," *Journal of Algorithms*, vol. 21, no. 3, pp. 579 – 597, 1996. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0196677496900607

[25] Y. Wang, L. Wu, and S. Wang, "A fully-decentralized consensus-based admm approach for dc-opf with demand response," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2637–2647, 2017.

[26] A. Kargarian, J. Mohammadi, J. Guo, S. Chakrabarti, M. Barati, G. Hug, S. Kar, and R. Baldick, "Toward distributed/decentralized dc optimal power flow implementation in future electric power systems," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 2574–2594, 2018.

[27] P. Li, Z. Wu, K. Meng, G. Chen, and Z. Y. Dong, "Decentralized optimal reactive power dispatch of optimally partitioned distribution networks," *IEEE Access*, vol. 6, pp. 74 051–74 060, 2018.

[28] Q. Peng and S. H. Low, "Distributed optimal power flow algorithm for radial networks, i: Balanced single phase case," *IEEE Transactions on Smart Grid*, vol. 9, no. 1, pp. 111–121, 2018.

[29] W. Lu, M. Liu, S. Lin, and L. Li, "Incremental-oriented admm for distributed optimal power flow with discrete variables in distribution networks," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6320–6331, 2019.

[30] D. K. Molzahn, F. Dörfler, H. Sandberg, S. H. Low, S. Chakrabarti, R. Baldick, and J. Lavaei, "A survey of distributed optimization and control algorithms for electric power systems," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2941–2962, 2017.

[31] D. Shchetinin, T. T. De Rubira, and G. Hug, "On the construction of linear approximations of line flow constraints for ac optimal power flow," *IEEE Transactions on Power Systems*, vol. 34, no. 2, pp. 1182–1192, 2019.

[32] J. Yang, N. Zhang, C. Kang, and Q. Xia, "A state-independent linear power flow model with accurate estimation of voltage magnitude," *IEEE Transactions on Power Systems*, vol. 32, no. 5, pp. 3607–3617, Sep. 2017.

[33] H. Yuan, F. Li, Y. Wei, and J. Zhu, "Novel linearized power flow and linearized opf models for active distribution networks with application in distribution lmp," *IEEE Transactions on Smart Grid*, vol. 9, no. 1, pp. 438–448, Jan 2018.

[34] S. Wang, Q. Liu, and X. Ji, "A fast sensitivity method for determining line loss and node voltages in active distribution network," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 1148–1150, 2018.

[35] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction*. MITP, 2015. [Online]. Available: https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf

[36] M. Al-Saffar and P. Musilek, "Reinforcement learning-based distributed bess management for mitigating overvoltage issues in systems with high pv penetration," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 2980–2994, 2020.

[37] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton, "A survey of monte carlo tree search methods," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 4, no. 1, pp. 1–43, 2012.

[38] E. Cotilla-Sanchez, P. D. H. Hines, C. Barrows, S. Blumsack, and M. Patel, "Multi-attribute partitioning of power networks based on electrical distance," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4979–4987, 2013.

[39] S. Blumsack, P. Hines, M. Patel, C. Barrows, and E. C. Sanchez, "Defining power network zones from measures of electrical distance," in *2009 IEEE Power Energy Society General Meeting*, 2009, pp. 1–8.

[40] E. Planas, J. Andreu, J. I. Gárate, I. Martínez de Alegría, and E. Ibarra, "Ac and dc technology in microgrids: A review," *Renewable and Sustainable Energy Reviews*, vol. 43, pp. 726 – 749, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364032114010065

[41] E. Unamuno and J. A. Barrena, "Hybrid ac/dc microgrids—part i: Review and classification of topologies," *Renewable and Sustainable Energy Reviews*, vol. 52, pp. 1251 – 1259, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364032115008412

[42] ——, "Hybrid ac/dc microgrids—part ii: Review and classification of control strategies," *Renewable and Sustainable Energy Reviews*, vol. 52, pp. 1123 – 1134, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364032115008333

[43] Wei Ren, R. W. Beard, and E. M. Atkins, "A survey of consensus problems in multi-agent coordination," in *Proceedings of the 2005, American Control Conference, 2005.*, 2005, pp. 1859–1864 vol. 3.

[44] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *arXiv*, 2016.

[45] C. Zhou, B. Huang, and P. Fränti, "An advantage actor-critic algorithm for robotic motion planning in dense and dynamic scenarios," in *arXiv*, 2021.

[46] D. Lee and J. Lee, "Incremental receptive field weighted actor-critic," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 1, pp. 62–71, 2013.

[47] N. P. J. U. L. J. A. N. G. L. K. I. P. A. Vaswani, N. Shazeer, "Attention is all you need," pp. 5998–6008, 2017.

[48] S. Yoon, Y. Choi, J. Park, and S. Bahk, "Stackelberg-game-based demand response for at-home electric vehicle charging," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 6, pp. 4172–4184, 2016.

[49] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *2018 Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 2, 2018.

[50] J. KE, f. xiao, H. Yang, and J. Ye, "Learning to delay in ride-sourcing systems: a multi-agent deep reinforcement learning framework," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2020.

[51] W. Zheng, W. Wu, B. Zhang, H. Sun, and Y. Liu, "A fully distributed reactive power optimization and control method for active distribution networks," *IEEE Transactions on Smart Grid*, vol. 7, no. 2, pp. 1021–1033, 2016.

[52] B. Zeng, J. Feng, N. Liu, and Y. Liu, "Co-optimized parking lot placement and incentive design for promoting pev integration considering decision-dependent uncertainties," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 1863–1872, 2021.

[53] R. Li, W. Wang, and M. Xia, "Cooperative planning of active distribution system with renewable energy sources and energy storage systems," *IEEE Access*, vol. 6, pp. 5916–5926, 2018.

[54] Y. Wang, X. Lin, and M. Pedram, "A stackelberg game-based optimization framework of the smart grid with distributed pv power generations and data centers," *IEEE Transactions on Energy Conversion*, vol. 29, no. 4, pp. 978–987, 2014.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3075247, IEEE Access

MOHAMMED AL-SAFFAR (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Alberta, Edmonton, AB, Canada in 2020. This paper is related to his Ph.D. work. His research is concerned with the optimization and management of distributed energy resources in power distribution systems, with the use of artificial intelligence techniques to process and interpret the modelling results. Moreover, his research explores the feasibility of fully distributed architectures that can interact with each other using deep learning and their applicability in different industry settings.

PETR MUSILEK (Senior Member, IEEE) received the Ing. degree (with great Distinction) in electrical engineering, and the Ph.D. degree in cybernetics from the Military Academy in Brno, Czech Republic, in 1991 and 1995, respectively. In 1995, he was appointed the Head of the Computer Applications Group, Institute of Informatics, Military Medical Academy, Hradec Kralove, Czech Republic. From 1997 to 1999, he was a NATO Science Fellow with the Intelligent Systems Research Laboratory, University of Saskatchewan, Canada. In 1999, he joined the Department of Electrical and Computer Engineering, University of Alberta, Canada, where he is currently a Full Professor. Since 2016, he served as a Director of Computer Engineering and an Associate Chair (Undergraduate). Currently, he is an Associate Chair (Research and Planning). His research interests include artificial intelligence and energy systems. He developed a number of innovative solutions in the areas of renewable energy systems, smart grids, wireless sensor networks, and environmental monitoring and modeling.

• • •