Deep Reinforcement Learning Based Coalition Formation for Energy Trading in Smart Grid

Mohammad Sadeghi and Melike Erol-Kantarci, *Senior Member, IEEE*
School of Electrical Engineering and Computer Science
University of Ottawa, Ottawa, ON
{msade033, melike.erolkantarci}@uottawa.ca

2021 IEEE 4th 5G World Forum (5GWF) | 978-1-6654-4308-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/5GWF52925.2021.00042

***Abstract*—Peer-to-peer energy trading is a promising approach to better integrate renewable energy resources, reduce customer costs and increase the reliability of the smart grid by employing microgrids and allowing them to share their surplus energy with each other using 5G-enabled communications. However, the varying nature of the generation and the demand of each microgrid impose a dynamicity and uncertainty on the system. In this paper, we address the problem of minimizing cost in the coalitional microgrid communities considering the dynamic nature of the system. We propose a deep reinforcement learning approach that helps to minimize the total cost through forming efficient coalitions. The results show 16% to 30% improvement in terms of cost minimization compared to an existing Q-learning-based scheme and a conventional coalitional game theory (CG)-based approach from the literature, respectively.**

***Index Terms*—Energy trading, machine learning, micro-grid, smart grid**

I. Introduction

The power grid has evolved drastically over the last two decades thanks to advances in information and communications technology (ICT) technologies. Although the main elements of the smart grid have reached maturity, careful integration of these elements in the new dynamic heterogeneous environment has remained as an open issue. Artificial Intelligence (AI) or specifically machine learning techniques are promising candidates to address the current challenges across the smart grid and microgrid infrastructure [1]. In recent years, machine learning gained huge attraction due to its great capability to enable many applications in various areas such as computer vision, robotics, 5G and so on. Smart grid, and specifically management of microgrids, poses challenges that are among the research domains that are targeted to be addressed with the design and employment of machine learning techniques.

In the literature, several prior attempts have employed machine learning to resolve microgrid related control problems. In [2], dynamic demand response and distributed generation management methods are proposed for residential microgrid communities. In [3] and [4], multi-agent correlated Q-learning approaches are introduced consequently to address the decentralized energy management problem. In [5], the problem of resource allocation in the wireless network for microgrid communities is investigated using a reinforcement learning-based algorithm.

One opportunity that comes from employing multiple microgrids in a distribution area is that they can trade energy among themselves, and hence contribute to a transactive energy system that is enabled by 5G network. Energy trading among close-by microgrids allows taking advantage of surplus generation in one microgrid to supply a nearby microgrid that is experiencing energy shortage. However, energy trading among microgrids always imposes costs in different aspects. The cost can be the result of the power loss in the line and the transformer interconnection; or it can be the result of the situation where there is no direct link between the buyer and seller and transferring energy occurs through intermediate microgrids, incurring extra costs. Apparently, as the distance between microgrids increases the cost increases. To this end, microgrids will trade energy with close-by neighbors, and as a result, proposing a method to efficiently pair nearby microgrids to involve in a peer-to-peer energy transaction is crucial. The underlying communication system is equally important however this paper deals with the control of multiple microgrids and assumes a reliable 5G network to enable message exchange between entities. Energy trading can be modeled as a game theory problem. Specifically, coalitional game theory is one of the methods that can address energy trading problems as in [6].

Although, in the literature, energy trading problem in a microgrid is tackled with the coalitional game and reinforcement approaches individually. There are fewer research attempts to employ both at the same time. In this paper, we propose a deep reinforcement learning approach to form coalitions to effectively address uncertainty in generation and demand in energy trading problems among microgrids. We call this scheme as Deep Q-Learning Based Coalition Formation (DQN-CF). In DQN-CF, as the agents interact through the iterations of reinforcement learning, they update their estimations, finally reaching coalitions that minimize cost. Our results show 16% and 30% improvement in cost minimization when DQN-CF is compared with an existing Q-learning-based scheme and a conventional coalitional game theory

978-1-6654-4308-1/21/$31.00 ©2021 IEEE
DOI 10.1109/5GWF52925.2021.00042

(CG)-based approach, respectively.

The rest of this paper is organized as follows. In Section II, the related work is summarized. In Section III, the system model is described. In Section IV, we introduce our deep reinforcement learning-based coalitional game scheme. Numerical results are provided in Section V and finally, the conclusion is presented in Section VI.

II. Related work

Energy trading problems are investigated in the literature using optimization, game theory and most recently machine learning. In [7], the authors proposed a learning automata-based method to tackle the energy trading problem in a dynamic microgrids community. In [8] and [9], energy trading problem has been investigated using hotbooting Q-learning approach and deep Q-network based approaches respectively.

Besides, these more recent machine learning based solutions, coalitional game theory, in which players cooperate to maximize a shared payoff, has been employed in several studies. It is usually considered that microgrids can form coalitions for a specific time interval where some microgrids have surplus energy and are willing to supply energy to the others that need to export energy. In [6], an energy trading problem in the microgrid community is investigated to minimize the power loss that occurs due to the transferring energy over power lines. Furthermore, in [10], the authors examine the problem of coordinated operation of cooperative microgrids in a distributed fashion. In [11], a nucleolus-based solution is proposed to fairly distribute the payoff among microgrids for transactive energy management in microgrid communities locally. In [12], the authors proposed a coalitional-based energy trading wherein each coalition, an auction-based matching is employed to calculate the utility of the coalition, and then coalition formation technique is used to partition the microgrids into coalitions. In [13] a two-stage algorithm is proposed for the energy trading problem of microgrids. At the first stage, a coalition formation algorithm is used, and then in the second stage, a matching game is employed to manage the energy exchange inside each coalition. In our prior work [14], we employed a Bayesian coalitional game-theory based scheme to address the uncertainty imposed by the penetration of EVs in the microgrid community. In [15] to tackle the problem of coalition formation in microgrid communities under uncertainty, we introduced a Bayesian reinforcement-based technique that lets microgrid agents to form coalitions by learning from their past experiences. Considering the dynamic nature of generation and demand and to tackle the resulting uncertainty, in this paper, for the first time, we introduce the Deep reinforcement learning-based coalition formation solution for peer-to-peer energy trading problem with the aim to minimize the cost of energy trading.

III. System Model

In this paper, we consider a system of M interconnected microgrids (MGs), and all the MGs are connected to the main macrogrid. The MGs can participate in an energy transaction with other nearby MGs. The energy trading opportunity gives the MGs the chance to export their surplus energy generation or to import energy when they have unsatisfied loads. In each epoch of our considered system, some of MGs are sellers (exporters) of energy and others are buyers (importers) of energy. This work aims to efficiently group buyers and sellers in a way that minimizes the cost of energy transactions compared to buying or selling energy to the macrogrid. In a group, or coalition, of MGs, there can be more than one seller and more than one buyer. This scenario can be modeled as a game-theoretical problem. In the rest of this section, we define the system.

A. Cost Function

We represent the generated energy and the demand of m-th MG by g_m and d_m respectively. Therefore, the equivalent surplus/shortage can be defined as $q_m = g_m - d_m$. q_m shows the amount of energy that m-th MG is required to import from or export to the grid which makes MGs involve in an energy trading transaction. This value is changing in each iteration and imposes uncertainty to the system. Energy trading transaction among MGs occurs with some cost. In this section, we first formulate these costs and then define them as an objective function that needs to be minimized. To this end, we introduce two set of costs during each energy transactions: *operational costs* and *virtual costs*. The operational cost denotes the cost associated with power loss during the energy transaction which includes the power loss in the line, transformer loss from different voltage levels (high, medium or low) and maintenance costs. We can formulate the power loss during energy transaction between m-th and n-th MG as follows:

$$PL(E_{mn}) = R_{mn} d_{mn} \frac{{E_{mn}}^2}{U_m^2} + \rho E_{mn}, \qquad (1)$$

R_{mn}, d_{mn}, and E_{mn} represent the resistance of line per km, the distance in km, and the required trading power in energy transaction between m-th and n-th MGs. Additionally, U_m shows the voltage of the line and ρ denotes the fraction of transformer power loss at connections with macrogrid. $\rho = 0$ is in the interconnections of MGs . Since always some part of the traded energy will be lost in the line, E_{mn} should be equal to demanded energy plus the amount of power loss in the line as follows:

$$E_{mn} = q_n + PL(E_{mn}) \qquad (2)$$

In the following, we define the virtual costs. Virtual costs associate with the hidden cost such as costs of involving a set of intermediate MGs to transfer energy from seller MG to buyer MG when the two MGs are not directly connected. This is a practical assumption since having a direct link between all MGs is not feasible in real scenarios. To this end, we assume that only close-by MGs are directly connected. In addition, any unpredicted costs such as the possibility of energy shortage which can result in a blackout when MGs are in islanding mode can be estimated as part of the virtual costs. We assume that virtual costs grow as the distance between MG increases and we formulate it as:

$$S_V = w d_{mn} E_{mn} \tag{3}$$

w is a weighting factor that can be formulated as follows:

$$w = \{ \begin{array}{ll} w_s & m,n \neq 0 \; and \; d_{mn} \leq d_{tr} \\ w_l & m,n \neq 0 \; and \; d_{mn} > d_{tr}. \\ w_0 & m = 0 \; or \, n = 0 \end{array} \tag{4}$$

Here, $w = w_s$ denotes energy trading happens in a closer distance than threshold distance (d_{tr}), and $w = w_l$ means the two MGs (m,n) have more than d_{tr} distance in between [16]. This way, we assume that there is no direct link between further than a threshold MGs and as the result, the virtual cost of such energy trading will be higher. Consequently, we can define the total costs as follows:

$$S_{mn} = w d_{mn} E_{mn} + \delta PL(E_{mn}) \tag{5}$$

The objective of this work is to minimize the total costs as follows:

$$\begin{gathered} \min \sum_{m=0}^{M} \sum_{n=0}^{M} S_{mn} \\ s.t: \sum_{m=1}^{M} q_m + \sum_{m=0}^{M} \sum_{n=0}^{M} P(E_{mn}) = \sum_{n=0}^{M} E_{0n} - \sum_{m=0}^{M} E_{m0} \end{gathered} \tag{6}$$

Considering the above-mentioned objective function and criteria, we can conclude that MGs prefer to trade energy with their close neighbors rather than further ones or with macrogrid. Having energy trading in short distances can help reduce costs associated with distance and also there will be zero transformer loss in any energy trading between MGs. Therefore, forming a group of MGs that can trade energy among themselves can be a promising method to reduce cost. In this paper, we use the coalition formation methodology to divide MGs into such groups. In the following section, we introduce the coalition formation.

B. *Coalition formation*

We define coalitions as a group of members with a coalition leader l_C. We assume that we have C coalitions and consequently C leaders. The leader is responsible to approve new members for the coalitions. Also, all communications with members happen through leaders. We represent each coalition with a pair (C, v_C). All the members of coalition C cooperate to maximize the total profit of the coalition known as the coalition value v_C. We define the coalition value as the negative form of the cost. In this manner, we ensure the minimization of cost as the members of coalitions, which are MGs in our problem, attempt to maximize their coalition values. The coalition value of each coalition consists of the cost of energy trading among the coalition members plus the cost of the energy transaction with the macrogrid. Therefore, the coalition value can be defined as:

$$v_C = -\sum_{m=0}^{|C|} \sum_{n=0}^{|C|} S_{mn}, \tag{7}$$

Here, the number of coalition members of coalition C is denoted by $|C|$. We use the index 0 to consider energy transaction (import or export) with macrogrid. After forming the coalition, coalition leader schedules energy trading in a manner that minimizes the total cost of energy trading among coalition members. Hence, the coalition payoff can be defined as the maximum achievable coalition value as follows:

$$v_C^{max} = \max \{-\sum_{i=0}^{|C|} \sum_{j=0}^{|C|} S_{mn}\}. \tag{8}$$

IV. Deep Q-Learning Based Coalition Formation (DQN-CF)

In this section, we introduce our DQN-CF scheme, where each coalition leader uses DQN to decide whether to accept an MG to the coalition or not based on minimizing the total cost of the coalition. In the following first we introduce the conventional Q-learning method and then we extend it to present the proposed DQN-CF algorithm. We will later use the Q-learning method as a benchmark to evaluate the DQN-CF.

A. *Q-Learning based Coalition Formation (QN-CF)*

Q-learning is one of the popular reinforcement methods that has been widely used in literature. The tuples of Q-learning can be defined as agents l, state s, action a and reward function r. In the proposed scheme, we define each element as follows:

1) *Agents*: In this scheme, we consider the leaders to be the agents of the system. Since we have more than one agent that performs actions, we have a multi-agent scenario.

2) *Actions*: In each time step, one MG has the chance to propose to join a new coalition randomly. The corresponding leader decides to accept or reject the joining proposal.Therefore, action space of leader l is a binary variable a^l.

3) *States*: We define the state of the agent l in the system as $s^l = \{\hat{q_1}, \hat{q_2}, .., \hat{q_M}, P_{index}\}$. The $\hat{q_m}$ is the quantized total surplus energy or shortfall of energy of the m-th MG. The $\hat{q_m}$ is zero for non member MGs

except for the proposer and P_{index} represents the index of the proposer.

4. *Reward*: The aim of the reward design is to minimize the total cost. The reward function is considered to be in the negative form of the coalition cost as follows:

$$\begin{aligned} r^l &= v_C^{max} \\ &= \max\{-\sum_{i=0}^{|C|}\sum_{j=0}^{|C|} S_{mn}\}. \end{aligned} \tag{9}$$

Q-learning chooses actions that maximize the expected current and future rewards to reach a sub-optimal policy. We update Q-value in Q-learning considering the Bellman's equation as below:

$$Q_\pi(s,a) = Q_\pi(s,a) + \alpha(r(s,a) + \gamma V_\pi(s') - Q_\pi(s,a)) \tag{10}$$

where α and γ denote the learning rate, and discount factor that shows the significance of future rewards respectively. The ϵ-greedy method is used to guarantee the action exploration as in [17].

B. DQN-CF

Q-table is used regular Q-learning to keep the record of cumulative reward corresponding to each action and state pair a and s and then the agent decides to choose the next action according to the Q-values. Although this works work well for small Q-tables with limited action-state space, when the action-state space is large, the memory will be larger and consequently, time complexity increases dramatically. To this end, deep Q-learning has been proposed to address these problems which are employed in many studies where a neural network is used to estimate the Q-values [18]. Considering the Q-value in (9), as the system converges, the Q-value $Q(s,a)$ is equal to $r(s,a) + \gamma V_\pi(s')$. Therefore, the neural network chooses $r(s,a)+\gamma V_\pi(s')$ as the training target and consequently $Q(s,a)$ is the predicted result. Considering the training target, we can define the loss function as:

$$L(w) = E(r(s,a) + \gamma V(s',w) - Q(s,a,w)) \tag{11}$$

where w, s and s' are the weight of the neural network, the current state and the next state, respectively. In Q-learning, the Q-values vary dynamically, and consequently, the target values will change, which may result in an unstable output. Also, the training data should be discontinuous, while the state and action transitions are consecutive in Q-learning. To address these issues and improve the DQN, experience replay and target network have been proposed as two DQN improvement solutions [19].

In this work, the LSTM which is a subset of recurrent neural network (RNN) is used to estimate the Q-values [20]. LSTM is capable of capturing the long-term dependencies comparing to the conventional RNN network, and at the same time addresses the problem of vanishing gradient. Considering the block diagram of the LSTM network in Fig, 1, x_t and $h_t - 1$ are the input value at time t and the output value at time $t-1$ respectively. $c_t - 1$ represents cell state at time $t-1$. the current output is shown by h_t and c_t denotes the current cell state.

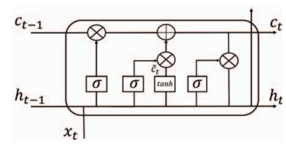

Fig. 1. Block diagram of the LSTM network.

The proposed algorithm is summarized in Algorithm 1. In every epoch of the system, one MG is selected at random to be the proposer. The proposer can choose randomly to stay in the current coalition or to join a new coalition. The coalition leader as the agent of the DQN will take action to accept or reject the joining proposal according to the estimated value function. Consecutive merges and splits happen until the system reaches a stable coalition formation in each estate.

Algorithm 1 Coalition formation with DQN for energy trading among MGs

```
Initialization: Initialize parameters α and γ.
At time t = 0:
for MG m = 1 to M do
        randomly select coalition C, set the power level q_i.
    Broadcast C to all MGs
end for
Main loop:
for Each time slot t = 1 to T do
    for leader l = 1 to L do
        update current coalition reward r_l(t)
        update the DQN estimatation using (10)
    end for
    DQN Coalition formation with probability of 1/M the proposer MG_i is selected from the set M and randomly choose coalition C;
    For the leader of coalition C:
    take an action a_l according to DQN estimation
    Sends a to the proposer m
If a_l = Yes then for all m ∈ C_k/{i} set m ∈ C_k and update the r_l.
end for
```

C. Baseline Algorithms

Baseline I - QN-CF: We introduced the Q-learning approach in section IV.A and we use this method to compare the proposed DQN based method with a well-known reinforcement learning based technique.

Baseline II - Conventional Coalitional Game Theory: The coalitional game theory based method is used as a benchmark as in [6].

V. RESULTS

In this work, MATLAB is used as simulation software. A region of 25kms by 25kms is considered in which MGs are assumed to be distributed randomly. The number of MGs varies between 6 and 18. and the number of coalition leaders varies between 3 to 7. A 24-hour interval is assumed for the simulation and load and generation patterns are generated randomly considering the Gaussian random variable. The same pattern with slight variations is repeated for each day periodically to simulate a longer time horizon [6]. We compare the proposed DQN-CF technique with Q-learning and CG-based methods. The results are achieved for 10 runs with 1500 iterations and the average results are reported. The simulation values are summarized in Table I. We use Adam optimizer in our DQN-CF method.

TABLE I
SUMMARY OF SIMULATION PARAMETERS

parameters	value
Line Resistance (R_{ij})	0.2
Medium Voltage (U_0)	50 kv
Low voltage (U_i)	22 kv
Transformer loss fraction (ρ)	0.02
Threshold distance (D_{tr})	5 km
Virtual cost parameter (w_s)	0.02
Virtual cost parameter (w_l)	0.04
Virtual cost parameter (w_0)	0.08
Scaling parameter (δ)	0.95
Learning rate of Q-learning (α)	0.5
Discount factor of Q-learning (γ)	0.8
Size of hidden layers	2
Number of hidden units	25
Training batch size	160
Size of replay memory	60
Training Interval	60

In Fig. 2, we present the average cost per user versus the number of MGs ranging from 6 to 18. As expected, by increasing the number of MGs, the cost will be reduced since MGs have more chance to make local coalitions in a dense network, resulting in less power transmission with a macrogrid. Moreover, DQN-CF demonstrates better results in terms of cost compared to the other algorithms. The proposed algorithm shows 4% to 16% improvement compared to CG and the sub-optimal QL-CF respectively.

In Fig. 3, to evaluate the effect of increasing power levels, we demonstrate the average cost per user versus the number of quantized power levels. As it is shown, when the power levels increase, the average cost decreases, as expected. As we increase the number of power levels, the quantization error will be reduced, and as a result, the performance of all the approaches improves while the proposed scheme performing significantly better than the others.

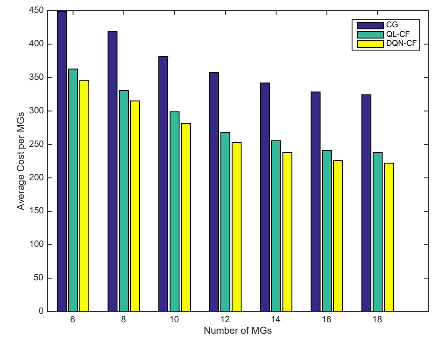

Fig. 2. Average cost versus number of MGs.

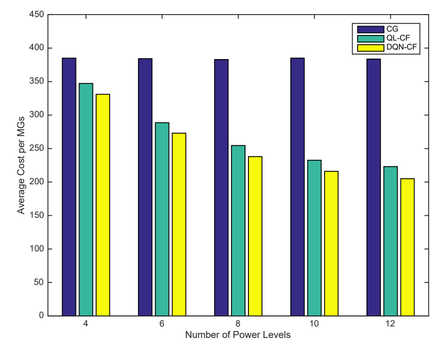

Fig. 3. Average cost versus number of power levels.

In Fig. 4, we present the average power loss per user versus the number of MGs. As the number of MGs increases, the power loss decreases. Moreover, since DQN-CF is designed to overcome the uncertainty, it demonstrates better results in terms of power loss compared to benchmark approaches up to 24% improvement with respect to conventional CG.

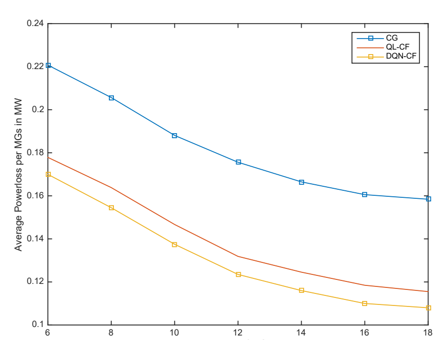

Fig. 4. Average power loss versus number of MGs.

In Fig 5, the average amount of energy transferred to the macrogrid versus the number of MGs is presented. As seen from the figure, DQN-CF causes less export to the macrogrid compared to the benchmark techniques. Additionally, as the number of MGs increases, there is more chance that nearby MGs join the same coalition since transferring energy with nearby MGs can result in less power loss and virtual costs, which also reduces the power exported to the macrogrid.

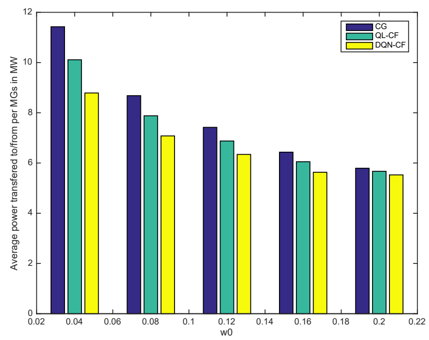

Fig. 5. Average energy transfer with Macrogrid versus number of MGs.

In Fig 6, the average cost per MG is plotted versus the number of iterations. Since we considered the cost as the reward, less value shows better performance. As it is expected, DQN-CF converges faster with better performance comparing to QL-CF.

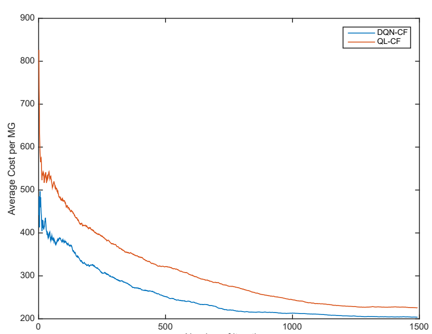

Fig. 6. Average cost per MG is plotted versus the number of iterations.

VI. Conclusion

Peer-to-peer energy trading and MG communities will play a vital role in future energy systems. In this study, we investigated the coalitional energy trading problem with the aim of cost minimization in a system with uncertainties. We proposed a deep reinforcement learning approach to overcome the uncertainties that arise from the power level of MGs which results in less energy transfer from macrogrid or distant MGs. We compared the proposed approach with Q-learning and CG schemes, and a significant reduction of (almost 16% and 30%) cost has been achieved.

References

[1] C. Jiang, H. Zhang, Y. Ren, Z. Han, K. C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Communications*, vol. 24, pp. 98–105, April 2017.

[2] B. Jiang and Y. Fei, "Dynamic residential demand response and distributed generation management in smart microgrid with hierarchical agents," *Energy Procedia*, vol. 12, pp. 76 – 90, 2011.

[3] H. Zhou and M. Erol-Kantarci, "Decentralized microgrid energy management: A multi-agent correlated q-learning approach," in *2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pp. 1–7, 2020.

[4] H. Zhou and M. Erol-Kantarci, "Correlated deep q-learning based microgrid energy management," in *2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pp. 1–6, 2020.

[5] M. Elsayed, M. Erol-Kantarci, B. Kantarci, L. Wu, and J. Li, "Low-latency communications for community resilience microgrids: A reinforcement learning approach," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1091–1099, 2020.

[6] W. Saad, Z. Han, and H. V. Poor, "Coalitional game theory for cooperative micro-grid distribution networks," in *IEEE International Conference on Communications Workshops*, pp. 1–5, 2011.

[7] S. Misra, P. V. Krishna, V. Saritha, and M. S. Obaidat, "Learning automata as a utility for power management in smart grids," *IEEE Communications Magazine*, vol. 51, pp. 98–104, January 2013.

[8] X. Xiao *et al.*, "Energy trading game for microgrids using reinforcement learning," in *Game Theory for Networks*, pp. 131–140, Springer International Publishing, 2017.

[9] L. Xiao *et al.*, "Reinforcement learning-based energy trading for microgrids [online]." https://arxiv.org/abs/1801.06285.

[10] C. Feng, F. Wen, S. You, Z. Li, F. Shahnia, and M. Shahidehpour, "Coalitional game based transactive energy management in local energy communities," *IEEE Trans. on Power Systems*, 2019.

[11] R. Lahon, C. P. Gupta, and E. Fernandez, "Coalition formation strategies for cooperative operation of multiple microgrids," *IET Generation, Transmission Distribution*, vol. 13, no. 16, pp. 3661–3672, 2019.

[12] J. Mei, C. Chen, J. Wang, and J. L. Kirtley, "Coalitional game theory based local power exchange algorithm for networked microgrids," *Applied Energy*, vol. 239, pp. 133 – 141, 2019.

[13] C. Essayeh, M. R. El Fenni, and H. Dahmouni, "Optimization of energy exchange in microgrid networks: A coalition formation approach," *Protection and Control of Modern Power Systems*, vol. 4, no. 1, p. 24, 2019.

[14] M. Sadeghi, S. Mollahasani, and M. Erol-Kantarci, "Power loss-aware transactive microgrid coalitions under uncertainty," *MDPI Energies*, p. to appear, 2020.

[15] M. Sadeghi and M. Erol-Kantarci, "Power loss minimization in microgrids using bayesian reinforcement learning with coalition formation," in *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1–6, Sep. 2019.

[16] M. Sadeghi, S. Mollahasani, and M. Erol-Kantarci, "Cost-aware dynamic bayesian coalitional game for energy trading among microgrids," in *IEEE International Conference on Communications Workshops (ICC)*, pp. 1–6, June 2021.

[17] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[18] M. Elsayed and M. Erol-Kantarci, "Ai-enabled future wireless networks: Challenges, opportunities, and open issues," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 70–77, 2019.

[19] V. Mnih and e. a. Kavukcuoglu, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.