Energy Management for Microgrids Using a Hierarchical Game-Machine Learning Algorithm

Rui Hu
School of Electrical and Computer Engineering
University of Pittsburgh
Pittsburgh, U.S
ruh14@pitt.edu

Alexis Kwasinski
School of Electrical and Computer Engineering
University of Pittsburgh
Pittsburgh, U.S
akwasins@pitt.edu

***Abstract*—This paper presents an energy management strategy for microgrids using a multiagent game-learning algorithm. This microgrid is powered by photovoltaic (PV) systems equipped with batteries and is intended to be operating in islanded mode. The proposed energy management strategy is applied to wireless communication networks by addressing the tradeoff between the communication signal's quality of service (QoS) and energy availability. A two-layer algorithm combining multiagent-game and reinforcement learning (RL) is designed for base stations (BSs) in order to accomplish the goal mentioned above. The proposed method shows improvement in the microgrid's performance and has a higher converging speed compared to a direct RL approach. The designed energy management algorithm was tested in multiple case studies.**

Keywords—microgrid, reinforcement learning, game theory, distributed control, multi-agent system, optimization, communication system, load shaping

I. INTRODUCTION

A microgrid is a local independently controlled electric system with distributed energy resources (DER), combined with storage devices and flexible loads. Such a system can be operated in a non-autonomous way or an autonomous way, depending on whether it is connected to the main grid [1]. Thanks to its independent control and local DERs, a microgrid can power its local area when the main grid is interrupted during a natural disaster [2]. This feature has drawn much attention from designers seeking high system resiliency, such as those in the communications industry. Considering the crucial role played by communication networks, a few companies have already explored using renewable energy sources and microgrids to power communication facilities consisting of a group of base stations (BSs). A practical question arises as of how to utilize the stored energy in such microgrid considering renewable resources' partial stochastic characteristics. Past research has been focusing on searching for an energy management strategy that solves this problem. One of the approaches is by switching on/off BSs and minimize the total load demand, as discussed in [3, 4] and [5]. Other methodologies considering green energy availability and delay performance include GALA [6], IDEA [7], and TEA [8]. These approaches rely on a central controller to search and broadcast the energy management strategy. The alternative way is to distribute the decision-making process to the individual BS controllers. Such an approach formulated the original energy management problem as a multiagent optimization problem and aimed at solving for an equilibrium. In [9], energy management is modeled as a multi-player game. However, the computation cost for a large scale system might be too costly and impractical [10]. A reinforcement machine learning algorithm was introduced in [11] so that the computation time is manageable. This reinforcement learning (RL) algorithm enables the BS to search for an operating point via trial and error but requires a few days of "training."

In this paper, we propose a hierarchical load response algorithm combing a virtual two-player game and RL process. Applying this algorithm, the controllers in a microgrid solve the immediate load response as a two-player game and adjust the user's load model using an RL algorithm. Because a two-player game could be solved in polynomial time, the controllers obtain a reasonable load plan fast in real-time and gain the capability of adapting to the unknown environment with the aid of RL. Additionally, compared to a direct RL approach, the two-player game simplifies the action searching space; thus, the converging speed is higher. As the study results will show, this approach obtains an energy management strategy with performance compared to an exhaustive heuristic search and has a shorter training period compared to direct RL algorithm.

II. MICROGRID ENERGY MANAGEMENT

An example of the proposed microgrid for communication networks is shown in Fig.1. The microgrid consists of a photovoltaic (PV) power generator, three communication BSs, and battery units. The batteries are responsible for absorbing excess generated power or powering the load when the generated power is insufficient. The BSs in this microgrid utilize communication traffic shaping (CTS) to adjust their energy consumption [12]. In a general form, the load at each BS could be expressed as

This work is supported and funded by the Hillman Foundation.

978-1-7281-4911-0/19/$31.00 ©2019 IEEE

$$P_{BS_i} = rt_i(P_b + P_c\delta(t)) = rt_i P_L \quad (1)$$

where rt_i is the ratio that the BS's load takes in the total microgrid load, P_b is the base BS power and P_c is the controllable power as a linear function of the traffic shaping factor (TSF) δ, and P_L is the microgrid total load demand. Generally, the higher δ is, the better the quality of the communication is. A traffic shaper controls ("shapes") the actual throughput (equivalent to the total volume of traffic) at the output of a communication system. Since the action of shaping traffic entails a reduction of bit rate, it will lead to an increased delay for data traffic or higher compression ratio for interactive video or speech traffic. In an LTE base station, a radio frame is divided into minimum units of transmit resources called "resource blocks" (RB). Without traffic shaping, all ongoing calls will require $R_B{}^T$ resource blocks. However, when applying traffic shaping, the actual number of active resource blocks becomes

$$R_B(t) \le \delta(t) R_B{}^T \quad (2)$$

Correspondingly, the maximum transmitted bit rate in this BS is limited based on the TSF by (2). Moreover, the impact of the quality of service (QoS) caused by choice of TSF is measured through the metric of peak signal-to-noise ratio (PSNR). A detailed discussion of the relation between QoS and PSNR could be found in [13]. Also, as shown in [14], PSNR could be approximated by

$$q_v = a \cdot \log(\delta \cdot r) + b \quad (3)$$

where q_v is the video quality measured in PSNR, r is the nominal bit rate, and *a* and *b* are constants based on the choice of source codecs. In this study, the codec type applied is H.265, and the corresponding values of *a* and b are 10.4 and -23.8. Details on how parameters in (3) values are obtained could be found in [13].

During operation, the actual load consumption depends on real-time variables such as communication traffic type (voice call/video stream) and the number of users. Additionally, the power generated by renewable resources, such as from photovoltaic (PV) modules and wind turbines, are partially stochastic [15]. Suppose their distribution functions are

$$P_L \sim X_L(t, \theta_1, \theta_2, ..., \theta_n) \quad (4)$$

$$G_R \sim X_R(t, \theta_1', \theta_2', ..., \theta_n') \quad (5)$$

where G_R is the power generated by a renewable source, t is time, θ_i and θ_i' are environmental conditions, and X_L and X_R are the distributions that the BS load and power follow. Correspondingly, the battery output power is also a random variable depending on the distribution of load and power generation

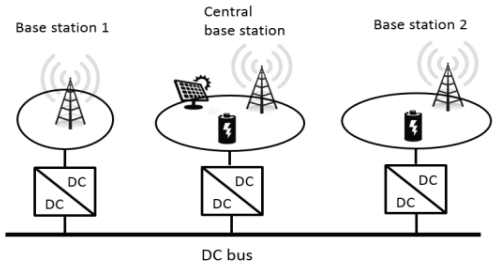

Fig. 1. Communication microgrid scheme.

$$P_B = G_R - P_L \quad (6)$$

Hence the battery energy output is

$$\Delta E_{bat}(t) = E_R - E_L = \int_{t_0}^{t} P_B(x)dx \quad (7)$$

where t_0 is the current time, E_r and E_L are the integrals of P_L and G_R overtime. Therefore, the battery state of charge (SoC) at time *t*

$$SoC(t) = \frac{E_{bat}(t_0) + \Delta E_{bat}(t)}{E_{bat\_full}} \quad (8)$$

is also a random variable, where $E_{bat}(t_0)$ is the battery's SoC at time t_0, and E_{bat_full} is the battery energy when fully charged. Suppose the cumulative density functions (CDF) of E_R is F_R, and the probability density function (PDF) of load consumption E_L is f_L, the pdf of battery energy ΔE_{bat} is then

$$f_{E_{bat}}(\Delta E_{bat}) = \int_{E_{BS=-\infty}}^{\infty} F_R(E_{bat} + E_L) f_L(E_L) dE_L \quad (9)$$

Hence the CDF of battery SoC(t)

$$F_{SoC}(SoC(t)) = \int_{-\infty}^{(SoC(t)-SoC_{now})E_{bat\_full}} f_{E_{bat}}(\Delta E_{bat}) d\Delta E_{bat} \quad (10)$$

where SoC_{now} is the battery's current SoC level. So the probability that the battery SoC is higher than any given value at time *t* is

$$P(SoC(t) > x) = 1 - F_{SoC}(x) \quad (11)$$

The SoC distribution information is essential for the users to evaluate its objective. A BS in the microgrid has multiple objectives: meet the load requirements, provide reasonable service quality and maintain sufficient stored energy. The last requirement comes from the fact again that the battery units are responsible for stabilizing the system bus voltage. If the batteries is fully discharged, the bus voltage in the microgrid could move away from the intended operating point and

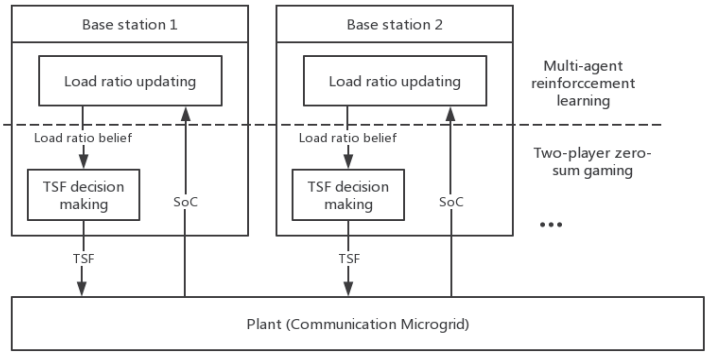

Fig. 2. Game-learning algorithm scheme.

cause the microgrid to shut down. Therefore, it is a good practice to have the microgrid batteries' SoC maintained at a relatively high level in island mode. In this project, the performance of the microgrid is evaluated by an objective function measuring the weighted sum of communication quality and battery SoC distribution:

$$obj(t) = w_{com} \cdot f_{com}(P_{BS,}t) + w_{SoC} \cdot P(SoC(t_d) > SoC_{goal}) \quad (12)$$

where $f_{com}(P_{BS,}t)$ calculates the average normalized peak signal-to-noise ratio (PSNR) of the communication network computed by (3). The BS's goal is to search for a TSF strategy $\delta(t)$ that maximizes the objective function (12). Details on how $f_{com}(P_{BS,}t)$ is calculated could be found in [9].

III. Game-Learning Algorithm

The proposed algorithm consists of two parts: the immediate TSF decision game and the BS load-ratio learning. The scheme of the algorithm is shown in Fig. 2. In this section, the details of the two parts are explained.

A. Two-player TSF game

Initially, the load demand (1) could be computed using the load and weather forecast information and modeled as a virtual two-player game similar to [9]. For a user with a load ratio rt_i, it assumes a virtual user takes all the rest of the load $(1-rt_i)(P_b + P_c\delta(t))$. Therefore, the objective function (12) depends on the TSF choices of both the actual and virtual user is

$$obj(t, \alpha_{-i}(t), \alpha_{-i}(t)) = w_{load} f_{com}(P_{BS}, t, \delta_i(t), \delta_{-i}(t)) + w_{SoC} \cdot P(SoC(t_d) > SoC_{goal} | \delta_i(t), \delta_{-i}(t)) \quad (13)$$

	Player II	P21	P22
Player I		δ_21	δ_22
P11	δ_11	$obj(t, \delta 11, \delta 21)$	$obj(t, \delta 11, \delta 22)$
P12	δ_12	$obj(t, \delta 12, \delta 21)$	$obj(t, \delta 12, \delta 22)$

Fig. 3. Example payoff table of a user.

where δ_{-i} indicates the virtual user's TSF action. Each user's objective is to maximize (13) considering the virtual player's possible action.

If a natural disaster hits the microgrid, the communications between BSs might be cut off, and the components could be damaged. In this condition, the behavior modes of BSs could be different. Some BSs may experience a surge in communication load demand, while other BSs may reduce their load consumption in order to conserve more stored energy. Therefore, the BSs in this microgrid do not necessarily share a common goal. Instead, it is safer for the BSs to assume the worst scenario, where its virtual player is trying to minimize their objectives. With this assumption, the objective of the BSs becomes

$$\max_{\delta_i} \min_{\delta_{-i}} obj(t, \delta_i(t), \delta_{-i}(t)) \quad (14)$$

which makes it a zero-sum game. The solution of (14) is the same as the solution of a corresponding linear programming problem

$$\begin{aligned} &Maximize \ \ z \\ &Subject \ to: \ \ A'x \geq z \cdot e \\ &e'x - 1 = 0 \\ &x \geq 0 \end{aligned} \quad (15)$$

where x is player I's strategy vector indicating its probability playing TSF actions, e is a vector of ones with the same length of x, A is the payoff table computed by deducing players' choice of TSF as shown in Fig. 3, z is the BS's expected payoff and pr_i is his probability of choosing the ith TSF. This correspondence of zero-sum game and linear programming is based on the connection between the Minmax Theorem and the Duality Theorem [16]. The linear programming problem could be solved by applying a simplex or interior-point algorithm [17].

B. Load-ratio learning

During the learning process, the BSs search for their optimal load-ratio policies through interactions with their environment and adapt their decision-making process in a

trial-and-error manner. At first, each BS is given a load-ratio list

$$L_r = [p_1, p_2, ..., p_M] \tag{16}$$

where p_i indicates the probability that the BS's load takes $\frac{i}{M}$ of the microgrid's total load. At time t, all the BSs randomly pick its load ratio according to their load-ratio lists:

$$rt_i(t) = \frac{i}{M} with\ \ probability\ \ \ p_i \tag{17}$$

After making the load-ratio choice, the BS conducts the two-player game solving and observe the resulting system status change. Then, a reward is computed and given to each agent to update its load-ratio policy, and the above process repeats. This process is similar to that of a policy iteration, but the updating law and its converging objective are different due to the multi-agent arrangement [18]. The learning sequence of the agent is shown below:

1. At time *t*, the BS chooses a load-ratio according to its load-ratio policy Lr_i. Suppose the load ratio taken is rt_i.
2. After conducting the two-player game and solving for its solution, each BS applies the obtained TSF strategy vector $\boldsymbol{x}$.
3. At the next LCI decision time $t+1$, the BSs collect the system status (SoC, PSNR information) and compute their payoffs. Suppose the reward of player *i* is $r_i(t)$.
4. BS updates its load-ratio policy according to the rule

$$L_{r_i}(t+1) = L_{r_i}(t) + \beta \cdot r_i(t)\left(e_{a_i} - L_{r_i}(t)\right) \tag{18}$$

where $0 < \beta < 1$ is a learning rate parameter and $\boldsymbol{e}_{a_i}$ is a unit vector with its rt_i th component unity. This algorithm is known as Linear Reward-Inaction algorithm L_{R-I}. One critical feature of this learning algorithm is that the convergence to pure NE of agents' strategies is guaranteed if the learning rate is sufficiently small, regardless of the number of players. By which it means that all probability vectors L_{r_i} converge to unity vectors. Such convergence is in terms of probability, and the mixed strategies NEs are not proven to be stable. However, compared to other multi-agent approaches such as mixed game and multi-agent Q-learning, this algorithm is one of the few that guarantees a form of convergence [19]. A full description of the L_{R-I} algorithm could be found in [20]. A modification made to this algorithm in this study is a set of limits to the load-ratio policy such that

$$p_i \le p_{\max} = 0.8 \forall i \tag{19}$$

$$p_i \ge p_{\min} = \frac{0.8}{M-1} \forall i \tag{20}$$

These limits are set so that the agents are not trapped in local optimums as the environment changes.

C. Design of reward function

The reward function has a critical influence on the BS's strategy obtained through RL. For example, if one action always leads to the highest payoff, the BS would develop a strategy to play that action only. In this study, two sets of reward functions are given to the BSs depending on their available information.

1) Reward function with communication

If the communication network in the microgrid is functioning normal and no component is lost, the BSs could exchange their choices of TSFs and battery SoC status with each other. So the original objective function (12) is commutable to all BSs; thus, (12) works as the reward function:

$$r_i = w_{com} \cdot f_{com}(P_{BS}, t, \delta_1(t), ..., \delta_N(t)) + w_{SoC} \cdot p(SoC(t_d) > SoC_{goal}) \tag{21}$$

A penalty is given to the BS when the probability of reaching the SoC goal is too small, as shown in (22). This setting encourages the BS to pick a lower TSF so that more energy is saved when the energy situation is critical.

$$r_i = 1 - \delta_i(t), P(SoC(t_{end}) \ge SoC_{gol}) < 0.5 \tag{22}$$

2) Reward function with local information

If the communication link in the microgrid is damaged or malfunctioning, it is possible that some of the BSs are unable to observe other BSs' moves and power-load profiles. In this situation, the disconnected BSs are limited to evaluate its performance with local information using the function shown in (23)

$$r_i = w_{load} \frac{1}{1+e^{-\delta_i(t)SoC(t)}} + w_{SoC} \frac{1}{1+e^{-\frac{SoC(t)}{\delta_i(t)}}} \tag{23}$$

which is an approximation of the original objective function, where a_i and SoC are the user's local TSF and battery SoC level. The first term $\frac{1}{1+e^{-\alpha_i SoC(t)}}$ is a normalized communication quality index, which gives the agent a high reward when both the SoC and $\delta_i(t)$ are large. The second term $\frac{1}{1+e^{-\frac{SoC(t)}{\delta_i(t)}}}$ is an approximated energy availability index. Similar to the standard operation reward, a penalty is given when the SoC level is critical

$$r_i = 1 - \delta_i(t), SoC(t) \le SoC_{\min} \tag{24}$$

IV. Analysis Verification

A case study applying this game-learning algorithm was conducted in MATLAB. The scheme of the simulated system is coincident with the one in Fig. 1, which consists of three BSs. It is assumed the users share the same fundamental parameters as listed in Table I. The users are asked to pick TSFs and update their load-ratio policy every hour. The initial load-ratio policy of a BS is shown in Table II, and the available TSF of a BS is $\delta(t)=[0.2,0.4,0.6,0.8,1.0]$.

Table I. Evaluation Parameter Values

Symbol	Parameter	Value
w_{com}	Communication quality weight	0.5
w_{SoC}	Energy availability weight	0.5
E_{bat}	Battery fully charged energy	24 kWh
$\overline{E_{P_{solar}}}$	Solar power expectation	1 kW
$\overline{E_{P_B}}$	BS base load expectation	0.2 kW
$\overline{E_{P_T}}$	BS traffic depended load expectation	0.8 kW
$\overline{V_{P_{solar}}}$	Solar power variance	4,000
$\overline{V_{P_T}}$	BS traffic depended load variance	4,000
SoC_{goal}	Desired battery SoC level	0.8
SoC_0	Initial Battery SoC level	0.7
BW	BS total bandwidth	10MHz
a	PSNR-rate bit curve parameter	10.4
b	PSNR-rate bit curve parameter	-23.8
r	Nominal transmit rate bit	2 Mbps

The PSNR and battery SoC of the overall system are demonstrated, and the system performance is compared to that of a direct RL algorithm and a heuristic algorithm with complete information. The exhaustive heuristic algorithm maximizes the objective function at each TSF decision time:

$$Maximize\ obj(t,\alpha(t))\forall t \tag{25}$$

The day-averaged cumulated objective function evaluates the performance of a microgrid

$$eva=\frac{\sum_{t=1}^{end} obj(t)}{day} \tag{26}$$

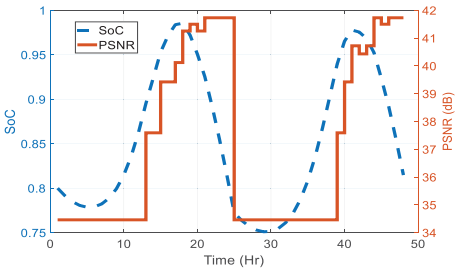

Fig. 4. PSNR and SoC of BS microgrid applying game-learning algorithm, normal condition.

Table I I. Initial Load-Ratio Policy of a User

rt_i	0.2	0.4	0.6	0.8	1
p_i	0.1	0.6	0.1	0.1	0.1

because such value reflects the system's long-term behavior in term of the objective function. This value is called the system performance index in the following context.

In the simulation, the system is operating normally where BSs are free to communicate with each other. The system PSNR and battery SoC during a two-day simulation is shown in Fig. 4. As the result shows, the obtained system PSNR is highly related to the system SoC level, which is similar to the one obtained by the heuristic algorithm (25) as shown in Fig. 5. Both algorithms ensure a system SoC over their desired goals (0.8), but the game-learning algorithm has a higher minimal PSNR. According to [14], a moderately good target for quality of video stream is 37 dB PSNR, whereas a 32 dB PSNR is considered as acceptable. Also, the overall system performance computed by (26) of the game-learning algorithm is $eva_{gl}=18.5058$ compared to that of heuristic one $eva_{heu}=18.9941$. Therefore, in the sense of overall objective function, the game-learning algorithm has a similar performance compared to the exhaustive heuristic algorithm.

A comparison between the game-learning and a direct RL approach is also conducted under the normal condition. If the TSF strategy is found through a direct RL approach linking SoC states with the TSF actions, such as the one in [11], the system PSNR and SoC would have a relationship as shown in Fig. 6 with a system performance $eva_{rl}=10.8693$, lower than that of the game-learning algorithm. This performance lost is mainly due to its exclusion of the time dimension in its searching space. Additionally, the direct RL approach demands a longer training period. With a learning rate of 0.1, the direct RL requires approximate 15 days until the system performance reaches a stable level, which could be observed from the training curve shown in Fig. 7. In comparison, when the same system is operated applying the game-learning

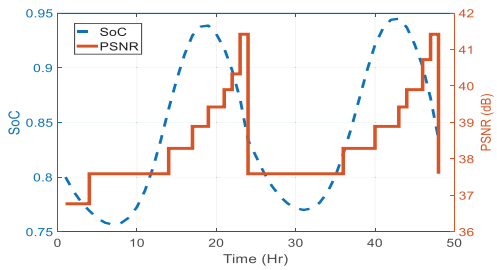

Fig. 5. LSR and system SoC applying heuristic algorithm, normal condition.

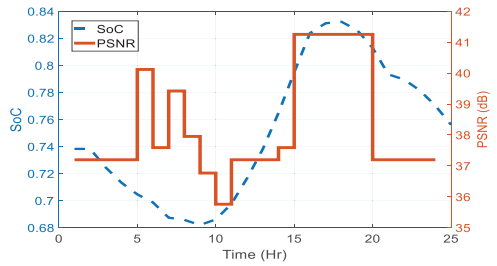

Fig. 6. LSR and system SoC applying direct RL, normal condition.

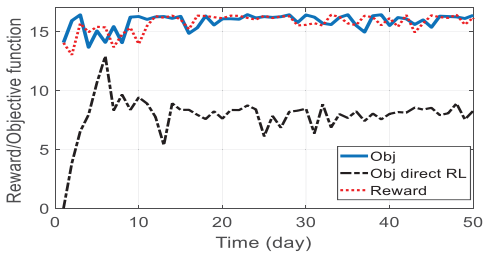

Fig. 8. Learning curve of a micogrid applying game-learning algorithm.

method, the obtained learning curve is shown in Fig. 8, which shows the system starts with a higher performance index as well as a faster-converging speed. The performance improvement is benefited from the game solving process, which takes consideration of power and load prediction. And the increased learning speed is likely caused by a smaller searching space: the agent applying direct RL has an SoC-TSF 2-dimension searching space while the agent in the game-learning algorithm needs only to explore the 1-dimension load-ratio policy.

V. Conclusion

This paper presented a two-layer game-machine learning energy management mechanism for microgrids to optimize its energy usage. In this particular case, the analysis focuses on a system applicable to wireless communication networks, but the same approach can be used in other applications with a partially controllable load. The simulation results show that the energy management strategy obtained by the game-learning algorithm has a better performance than solely applying the reinforcement learning and is close with an exhaustive heuristic search algorithm. Benefiting from the reduced searching space, converging speed of the proposed algorithm is higher than a direct reinforcement learning algorithm. Additionally, this algorithm showed strong resilience against system damage such as partial power loss and communication network failure. In the future, the performance of the adpating feature of the algorithm under dynamic environment will be valuated.

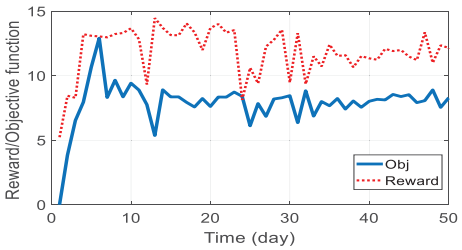

Fig. 7. Learning curve of a micogrid applying direct RL algorithm.

Reference

[1] N. Hatziargyriou, Microgrids : Architectures and Control. New York, United Kindom: John Wiley & Sons, Incorporated, 2013.

[2] M. Hanna, "When disasters strike distributed systems," vol. 15, ed. Newton: King Content Company, 1995, p. 54.

[3] E. Oh, K. Son, and B. Krishnamachari, "Dynamic Base Station Switching-On/Off Strategies for Green Cellular Networks," IEEE Transactions on Wireless Communications, vol. 12, no. 5, pp. 2126-2136, 2013.

[4] N. Yu, Y. Miao, L. Mu, H. Du, H. Huang, and X. Jia, "Minimizing Energy Cost by Dynamic Switching ON/OFF Base Stations in Cellular Networks," IEEE Transactions on Wireless Communications, vol. 15, no. 11, pp. 7457-7469, 2016.

[5] A. Stavridis, S. Narayanan, M. D. Renzo, L. Alonso, H. Haas, and C. Verikoukis, "A base station switching on-off algorithm using traditional MIMO and spatial modulation," in 2013 IEEE 18th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 2013, pp. 68-72.

[6] T. Han and N. Ansari, "Green-energy Aware and Latency Aware user associations in heterogeneous cellular networks," pp. 4946-4951: IEEE.

[7] D. Liu, Y. Chen, K. K. Chai, and T. Zhang, "Distributed delay-energy aware user association in 3-tier HetNets with hybrid energy sources," pp. 1109-1114: IEEE.

[8] V. Chamola, B. Krishnamachari, and B. Sikdar, "Green Energy and Delay Aware Downlink Power Control and User Association for Off-Grid Solar-Powered Base Stations," IEEE Systems Journal, vol. 12, no. 3, pp. 2622-2633, 2018.

[9] R. Hu, A. Kwasinski, and A. Kwasinski, "Mixed strategy load management strategy for wireless communication network micro grid," pp. 1-8: IEEE.

[10] C. D. a. C. H. Papadimitriou, Three-player games are hard. 2005.

[11] R. Hu and A. Kwasinski, "Energy management for microgrids using a reinforcement learning algorithm " in 2018 IEEE Green Energy and Smart Systems Conference (IGESSC), 2018.

[12] "Microgrids for disaster preparedness and recovery with electricity continuity plans and systems: with electricity continuity plans and systems," in Premium Official News, Newspaper Article, ed: Plus Media Solutions, 2015.

[13] A. Kwasinski and A. Kwasinski, "Integrating cross-layer LTE resources and energy management for increased powering of base stations from renewable energy," pp. 498-505: IFIP.

[14] A. Kwasinski and A. Kwasinski, "The role of multimedia source codecs in green cellular networks," vol. 2016-, pp. 1-6: IEEE.

[15] S. Vandael, B. Claessens, D. Ernst, T. Holvoet, and G. Deconinck, "Reinforcement Learning of Heuristic EV Fleet Charging in a Day-Ahead Electricity Market," IEEE Transactions on Smart Grid, vol. 6, no. 4, pp. 1795-1805, 2015.

[16] S. Homer and A. L. Selman, Computability and complexity theory, 2nd;2; ed. (no. Book, Whole). London; New York;: Springer, 2011.

[17] H. Karloff, Linear programming (Progress in theoretical computer science). Boston: Birkhäuser, 1991, pp. viii, 142 p.

[18] R. S. Sutton, A. G. Barto, and I. netLibrary, Reinforcement learning: an introduction (no. Book, Whole). Cambridge, Mass: MIT Press, 1998.

[19] L. Buşoniu, R. Babuška, and B. De Schutter, "Multi-agent Reinforcement Learning: An Overview," vol. 310Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 183-221.

[20] P. S. Sastry, V. V. Phansalkar, and M. A. L. Thathachar, "Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information," IEEE Transactions on Systems, Man, and Cybernetics, vol. 24, no. 5, pp. 769-777, 1994.