Resiliency-Aware Power Management of Microgrids using Agent-based Dynamic Programming and Q-learning

Farshina Nazrul Shimim*, Mohammad Alali*, Hashem Nehrir*, John Sheppard†, Maryam Bahramipanah*,
and Zagros Shahooei*
Dept. of Electrical and Computer Engineering*, Gianforte School of Computing†,
Montana State University, Bozeman, MT

***Abstract*—Appropriate planning and optimization strategies for day-ahead power management play important roles in efficient operation of Microgrids (MGs). Due to the uncertainties in electricity demand and renewable generations, and the multi-objective (MO) nature of MG power management, conventional optimization techniques have not been as effective in giving satisfactory results. This paper aims at solving the day-ahead power management problem as a MO optimization problem, with a focus on increasing the system's resiliency using an agent-based Dynamic Programming (DP) approach named Value Iteration (VI) and a model-free Q-learning (QL) algorithm. The two objectives of the MO problem are: maximizing load serviceability and minimizing operational cost. Both the approaches are data-driven, and the behavior of the agent of each component of a MG is formulated as a finite-horizon Markov Decision Process (MDP). VI guarantees an optimal solution to the MO problem given the MDP model, and QL has the ability to work under uncertainty and incomplete information. The effectiveness of the two algorithms have been evaluated using a benchmark MG test system.**

***Index Terms*—Power Management, Multi-agent System, Markov Decision Process, Islanded Microgrid.**

I. Introduction

Dividing a distribution system into several independent and autonomous entities called Microgrids (MGs) has proven to be a good strategy to deal with the uncertainties of renewable energy resources, electricity demand, and market retail price [1], [2], [3].

Many studies have been conducted to resolve the issue of power management under these uncertainties in MGs. [4] gives an overview of the importance of centralized controllers for stability, power quality, protection and power management of the MG. Stochastic programming is explained in [5] as an approach for power management. Furthermore, in [6], a two stage stochastic programming was employed to minimize the cost due to the variable nature of renewable energy resources, and in [7], constrained stochastic programming is used to consider the MG limitations. Value Function Approximation is used in [8] to solve a bench-marked ESS management problem. Further, in [9], a deep Recurrent Neural Network is used to approximate the value function while considering power flow constraints.

In general, the agent-based power management in a MG on a day-ahead basis can be modeled as a sequential decision-making problem [9], [10]. Under this scheme, the agents controlling the different components of the MG reduce the overall cost while maximizing the utilization of its microsources for a finite-horizon (e.g., 24 hours), looking at the electricity demand. These utilities include Dispatchable Generators (DGs), Renewable Energy Sources (RESs), Energy Storage Systems (ESSs), and Demand Response (DR) on Price-Sensitive load (PSR). This paper addresses a MG where the output of the DGs, charging/discharging paradigm of the ESSs, and the PSR can be regulated based on different operating conditions with a primary focus on maximizing the load coverage and a secondary focus on minimizing the cost, using VI and QL, which is in the domain of Reinforcement Learning. Therefore, making the power management problem multi-objective.

II. Methodology

In this paper, control of the MG components is formulated as a Markov Decision Process (MDP). We address the load variations and possible duration and intensity of an upcoming extreme event, using its probability of occurrence (assumed to be known from weather forecast) in our simulation. At each discrete time-step (each hour, for example) the output powers for the controllable utilities (DGs, ESSs, DR) are considered as actions, whereas the state-space consists of the supplied electricity demand and the State-of-Charge (SoC) of the ESS. The reward function for each agent is actually a penalty when considering their operating cost or deviation from the target load (the higher the operating cost or deviation from the target load results in higher penalties). The agents interact with the environment to achieve their goal of maximizing their objective functions.

A. Microgrid Power Management

In general, a MG is connected to the main grid through a Point of Common Coupling (PCC). In our simulation studies, the proposed methods are applied on a modified benchmark MG test system (MG3 from [11]), as shown in Fig. 1. The components of the MG are: a Diesel Generator as DG, several solar photovoltaic (PV) systems, several ESSs (an ESS at each bus with a PV system, not shown in Fig. 1), and

This work was supported by the US National Science Foundation, under Award 1806184.

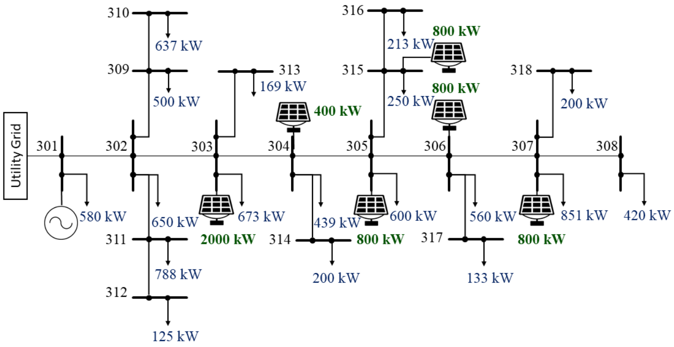

Fig. 1. Single-line diagram of the benchmark MG system.

some Controlable Loads (CLs). The day-ahead scheduling is performed for a 24-hour horizon with a 1-hour time-step. In this paper, the power management problem is formulated as a MDP for three agents: a DG agent for the operation and control of the dispatchable generators, a DR agent to perform demand response on the CLs, and an ESS agent to charge/discharge the ESSs. The total PV power is used when available. Therefore, no agent is used for PV generation.

B. Formulation of MDP

A finite-horizon MDP can be defined as a 5-tuple $\mathcal{M} =< \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, H >$ where, $\mathcal{S}$ is a finite set of state variables, $\mathcal{A}$ is a finite set of actions, $\mathcal{R}$ is the immediate reward observed after reaching a state, $s \in \mathcal{S}$. $\mathcal{T}$ is the set of transition probabilities for the transition of one state to another with a specific action [12], and H is the length of the horizon. For simplicity of implementation for both Value Iteration (VI) and Q-learning (QL) algorithms, the state-action space is discretized to their nearest integers. For the implementation of VI algorithm in our power application, the transition probability of reaching a particular state s_j using an action a_{ij} from state s_i is assumed to be 1, $s_i, s_j \in \mathcal{S}, a_{ij} \in \mathcal{A}$ [13]. The detailed models of the associated agents are described in the rest of this section.

1) DG agent: The DG agent is responsible for setting the output of the DG located in the MG, as per the electricity demand. In a MG, if P_{DG}^t is the active power output of the DG at a time step t, then for each DG, the operational constraints are as follows:

$$P_{DG}^{min} \leq P_{DG}^t \leq P_{DG}^{max} \tag{1}$$

$$|P_{DG}^t - P_{DG}^{t-1}| \leq \rho \times \Delta t \tag{2}$$

Here, the max and min terms represent the maximum and minimum output of the generator [9]. The term ρ represents the limit of the increase/decrease of generation at each time step, and is known as the Generation Rate Constant (GRC).

For the DG agent, the set of states consists of the possible load coverage by the DG, and is defined by $\mathcal{S}_{DG} = \{s_{DG_j}^t | \quad \forall j \in \mathcal{S}_{DG}, P_{DG}^{min} \leq s_{DG_j}^t \leq P_{DG}^{max}\}$. Only the active power outputs have been considered in this work. The action set is the range of feasible discrete set-points of the generator outputs, defined as: $\mathcal{A}_{DG} = \{a_{DG_i}^t | \quad \forall i \in \mathcal{A}_{DG}, |a_{DG_i}^t| \leq \rho \times \Delta t\}$

Each DG has an operational cost, which is a function of the active power output of the DG. For this paper, a quadratic cost function is considered, as defined in [14]. The cost function denotes that the higher output power results in higher cost, which is a penalty, denoted by a negative sign. At the same time, the DG focuses on maximum coverage of load. Therefore, the reward function for the DG agent is defined as follows:

$$\begin{aligned} R_{DG}(s_{DG_j}^t, a_{DG_i}^t) = &-\frac{(P_L^t - P_{PV}^t - P_{DG}^t)}{N_1} w_1 \\ &- \frac{\{(\alpha + \beta(P_{DG}^t) + \gamma(P_{DG}^t)^2) - (\lambda^t P_{DG}^t)\}}{N_2} w_2 \end{aligned} \tag{3}$$

where, α, β and γ are the generation cost coefficients [14], P_L^t and P_{PV}^t are the predictions of electricity demand and PV output, λ^t is the market electricity price forecast at time step t. w_1 and w_2 are the weight factors, and N_1 and N_2 are the normalization factors, respectively [15]. Here, it should be noted that during an extreme event resulting in grid blackout, which islands the MG, the term λ^t is zero.

2) DR agent: The task of the DR agent is to perform PSR on CLs. In case of insufficient generation, a percentage of the electricity demand is available for curtailment (the maximum percentage is given in Table I), associated with a penalty for causing discomfort to the user. The total electricity demand for the time horizon H is denoted by $\mathcal{P}_L = \{P_L^t, \quad \forall t = 1, 2, 3, ...H\}$. For the DR agent, only the active power is considered.

The state space for the DR agent is the permissible range of the curtailed loads $\mathcal{S}_{DR} = \{s_{DR_j}^t | \quad \forall j \in \mathcal{S}_{DR}, P_L^{t\,min} \leq s_{DR_j}^t \leq P_L^{t\,max}\}$. The action space for the DR agent is defined as $\mathcal{A}_{DR} = \{a_{DR_i}^t | \quad \forall i \in \mathcal{A}_{DR}, |a_{DR_i}^t| \leq P_{DR}^{t\,max}\}$. $\mathcal{A}_{DR}$ is restricted by the range of deviations of set-points, $P_{DR}^t = P_L^t - P_{sup}^t$. The load-shedding penalty for the DR is expressed as the following piece-wise linear model [9]:

$$P_{CL}^t = \begin{cases} \delta_0(P_L^t - P_{sup}^t) + \phi_0, & P_{L_{max}}^t \geq P_{sup}^t \geq P_{L_1}^t \\ \delta_1(P_L^t - P_{sup}^t) + \phi_1, & P_{L_1}^t \geq P_{sup}^t \geq P_{L_2}^t \\ \delta_2(P_L^t - P_{sup}^t) + \phi_2, & P_{L_2}^t \geq P_{sup}^t \geq P_{L_{min}}^t \end{cases} \tag{4}$$

Here, $\delta_0, \delta_1, \delta_2, \phi_0, \phi_1$, and ϕ_2 are the constant coefficients for the load-shedding penalty model. P_{sup}^t is the amount of load that is not curtailed, and hence to be supplied. Both $P_{L_1}^t$ and $P_{L_2}^t$ are set-points representing the mild, and moderate curtailment levels, as described in [16]. However, at the same time, the DR agent should consider the cost of supplying the loads, as per market electricity price. Therefore, the overall reward function for the DR agent becomes:

$$R_{DR}(s_{DR_j}^t, a_{DR_i}^t) = -\frac{P_{CL}^t}{N_3} w_3 - \frac{\lambda^t(P_L^t - P_{PV}^t)}{N_4} w_4. \tag{5}$$

Here, w_3 and w_4 are the weight factors, and N_3 and N_4 are the normalization factors, respectively [17].

3) ESS agent: The ESS agent is responsible for maximum load coverage during an extreme event, when the generation units are out of service. The agent considers the extreme event forecast at each hour of the next day, starts storing energy to supply the highest possible loads during the extreme event. Prior to the extreme event, additional energy will charge the ESS. For controlling the amount of energy stored in the ESS, the State of Charge (SoC) of the battery at each time step is defined as follows:

$$SoC^t_{ESS} = \frac{E^t}{E_{Cap}} \tag{6}$$

where E^t is the amount of energy stored in the ESS at time step t, and E_{Cap} is the energy capacity of the storage unit.

Typically, ESS should not be fully discharged; otherwise the storage unit might be seriously damaged. The limits for SoC which is known as capacity constraint of the ESS is shown below:

$$SoC^{min}_{ESS} \leq SoC^t_{ESS} \leq SoC^{max}_{ESS} \tag{7}$$

where SoC^{min}_{ESS} and SoC^{max}_{ESS} respectively denote the minimum and maximum allowed SoC at each time step. Also, E_{max} and E_{min} indicate the maximum and minimum allowed stored energy.

Looking at Eq. 7 the state space for the ESS agent can be represented as $\mathcal{S}_{ESS} = \{s^t_{ESS_j}| \quad \forall j \in \mathcal{S}_{ESS}, SoC^{min}_{ESS} \leq s^t_{ESS_j} \leq SoC^{max}_{ESS}\}$ The charging and discharging rates of the ESS (P_{ESS}) are considered as actions for the ESS agent. Accordingly, the action space for this agent is described as $\mathcal{A}_{ESS} = \{a^t_{ESS_i}| \quad \forall i \in \mathcal{A}_{ESS}, |a^t_{ESS_i}| \leq P^{max}_{ESS}\}$.

The reward function of the ESS agent is:

$$R_{ESS} = \begin{cases} SoC^{max}_{ESS} - SoC^t_{ESS} & \mu^t > 0.1 \ \wedge \ \mu^t \neq 1 \\ SoC^t_{ESS} \times E_{Cap} - P^t_L & (\mu^t = 1 \ \vee \ M = 1) \\ SoC^t_{ESS} - SoC^{min}_{ESS} & \text{otherwise} \end{cases} \tag{8}$$

where μ^t represents the probability of the extreme event at time step t. M is a flag indicating whether the grid is still under maintenance, or it is healthy. When healthy, the MG can be reconnected to the grid.

C. Power Management in a MDP framework

The agent-based scheduling for a MG is formulated as an optimization problem under a MDP framework with the following set of objective function:

$$\max_{a_{DG}, a_{DR}, a_{ESS}} \sum_{t=1}^{H} \{R_{DG}, R_{DR}, R_{ESS}\} \tag{9}$$

subject to (1),(2),(4),(6),(7)

D. Solution Approaches for the MDP

To solve the above-mentioned MDP problem, the Value Iteration (VI) algorithm is applied [18]. In this DP method, the optimal value of a state is calculated with the help of the following equation [19]:

$$Q(s_t, a_t) := E[r|s_t, a_t] + \Gamma \sum_{s'_t \in \mathcal{S}} P(s'_t|s_t, a_t) V(s'_t) \tag{10}$$

where, $E[r|s_t, a_t]$ denotes the expected immediate reward. The probability of transition from state s_t to s'_t by taking an action a_t, is represented by $P(s'_t|s_t, a_t)$. Γ is the discount factor, which determines the impact of the future rewards for a specific state. The values are said to converge when the difference between the values obtained from two consecutive iterations are less than a pre-defined threshold, θ.

$$V(s_t) = \max_{a \in \mathcal{A}} Q(s_t, a_t) \tag{11}$$

$$\max_{s \in \mathcal{S}} |V(s_t)^k - V(s_t)^{k+1}| < \theta \tag{12}$$

The MDP has also been solved using a temporal difference algorithm, namely Q-learning (QL). This algorithm learns the best action to move from one state to another based on the maximum expected values of the reward functions, without knowing the transition probability. The transition probability is used in the case of VI, but not in the case of QL. The experience gathered by each agent from the interaction with the state-space is recorded in a look-up-table, named Q-table. Using a specific state-action pair, the agents interact with the environment and the corresponding Q-table values are updated by a running average mechanism [20], [21].

III. Numerical Experiments and Results

A. Test Case

The proposed method is validated on the benchmark test MG system as a grid-connected MG (MG3 from [11]), as shown in Fig. 1. A 24-hour time horizon is considered for the whole simulation with a 1-hour window time-step. In our simulation, an extreme event occurs at hour 12 when both the grid and the DG go out of service. The extreme event continues until hour 14, after which a maintenance phase starts for the grid and the DG until hour 21, as shown in Fig. 6. The probability of occurrence of an extreme event at each time step is also shown in the figure. Note that, the probability keeps increasing before hour 12 and starts decreasing when the extreme event is over at hour 14. The forecasted electricity demand, PV output data, and system specification is adopted from [11]. The retail electricity price data were obtained from [17]. The technical information about the components of the MG is displayed in Table 1.

For the DG agent, the GRC constant, ρ, the generator cost coefficients α, β and γ were selected as 700 kW, \$14.67, \$0.1709/kWh, and \$0.0001773/(kWh)^2, respectively. For the DR agent, the constant coefficients δ_0, δ_1 and δ_2 were set as

TABLE I
POWER/ENERGY CAPACITY OF THE COMPONENTS OF THE TEST SYSTEM

Components	**Max. Capacity (MW)**	**Min. Capacity (MW)**
DG	7	2.5
PV	5.6	0
P^t_L	7.988	6.63
CL	40% of P^t_L	0% of P^t_L
ESS	28.6 (MWh)	4.967 (MWh)

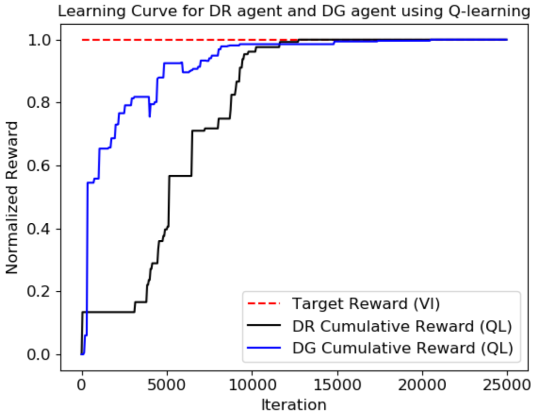

Fig. 2. Cumulative reward of the 24-hour planning horizon for the DG agent and the DR agent at each iteration using QL algorithm.

0.3, 0.5 and 0.75, respectively. All the ϕ values were set to 0. The curtailment levels, $P^t_{L_1}$ and $P^t_{L_2}$ were selected as 90% and 30% of the permissible range, $\mathcal{S}_{DR}$, respectively. The weight factors for the reward functions were set as, $w_1 = 0.7$, $w_2 = 0.3$, $w_3 = 0.7$ and $w_4 = 0.5$, to ensure that the main objective of the agents is maximizing load coverage rather than minimizing cost.

For VI, Γ and θ were chosen to be 0.99 and 0.001, respectively. For QL, the learning rate η is selected to be 0.1. The ϵ-*greedy* algorithm has been used for the QL action selection [19], and the value of ϵ (exploration rate) was initially chosen as 0.9 for every state-action pair in the Q-table with a decay rate of 0.95 for each visit of the agent.

B. Agent-based Power Management

The DG agent and the DR agent were trained to find a strategy for the DG, and the supplied load based on the retail price and customer discomfort level, respectively. Whereas, the ESS agent learns a strategy to prepare and provide energy to the MG to cover the essential loads (when possible) in case of a grid blackout due to a fault or an extreme event. During this period, the MG is islanded, and the DG is assumed to be out of service. Therefore, PV and ESS will cover the essential loads (assumed to be 50% of the total load). Fig. 2 shows the learning curve of the DG agent and DR agent for QL with respect to the number of iterations. The reward obtained through value-iteration is considered to be the target reward. From Fig. 2, the DR agent reaches the target value in about 12000 iterations and the DG agent in about 19000 iterations.

The selected states by the DG agent and the DR agent using both VI and QL are shown in Figs. 3 and 4, respectively along with the forecasted electricity demand and the renewable generation (PV). The policy that sequentially transits from one optimal state to another is considered to be the optimal policy.

The learned optimal policy for the ESS is demonstrated in Fig. 5. As shown in the figure, from $t = 3$ (when the extreme event probability is more than 10%) the battery starts storing as much energy as it can (SoC^{max}_{ESS}) (Fig. 6) so that it can cover

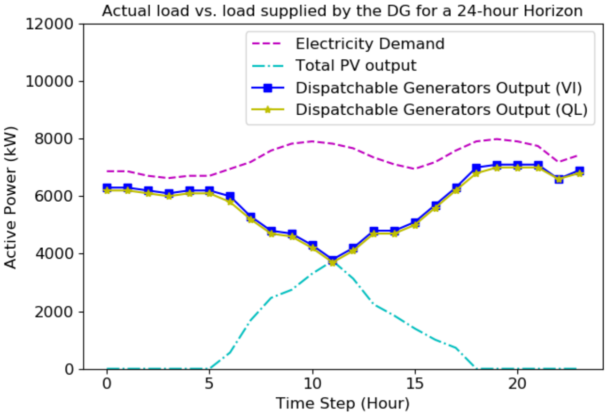

Fig. 3. Economic Dispatch of the DG agent using VI and QL.

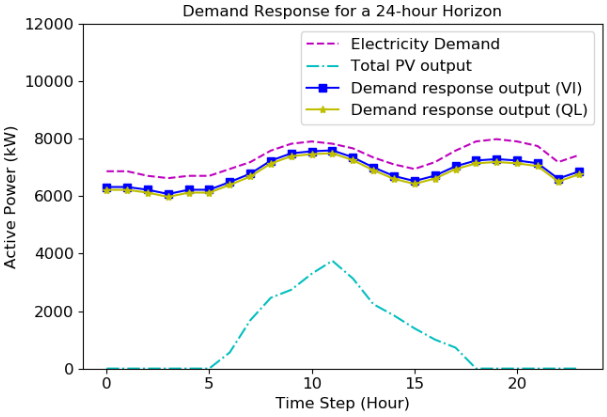

Fig. 4. Day-ahead scheduling of the DR agent using VI and QL.

maximum loads possible during the extreme event. After the recovery time, the normal operation of the MG is continued.

IV. Discussion

As mentioned in section II, the reward function for each agent has two inversely proportional components of different weights. Considering that the main focus of the agents is maximum load coverage, Figs. 3 and 4 show that using both the VI and QL algorithms, the agents learn to operate in between the feasible state spaces and take proper actions. For the DG agent, when the load change is more than the GRC constant (2), the change of generation is bound by the GRC. On the other hand, the DR agent is mostly operating within (80-100)% of the actual electricity demand at each hour. Finally, the performance of the system was examined in the presence of an ESS agent, in a scenario where there was a possibility of an extreme event. As shown in Fig. 5, before hour 6, 90% of the load is supplied by the generator. From hour 6 to 11, the combination of the generator and PV supplies the load. However, during the extreme event, since the generator is out of service, the combination of the PV and

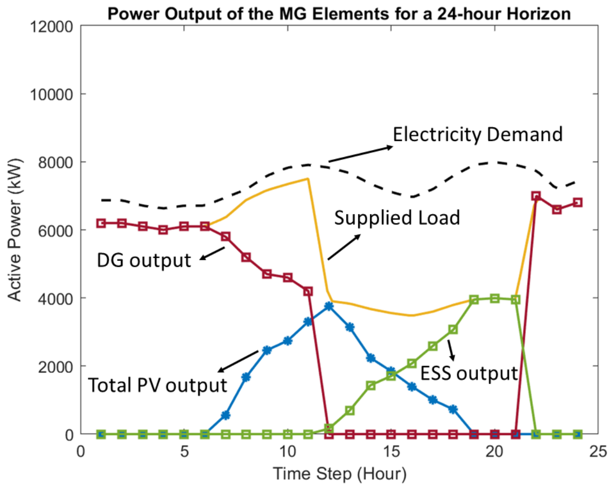

Fig. 5. Load coverage of the ESS agent under an extreme event scenario.

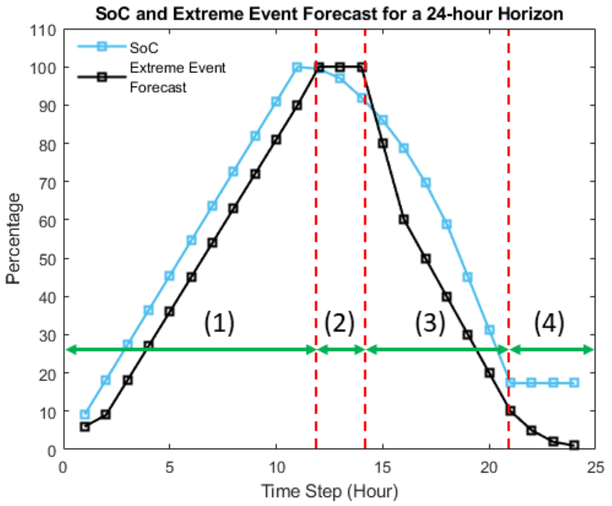

Fig. 6. Percentage of State of Charge (SOC) of the ESS for the 24-hour horizon, along with the probability of occurrence of an extreme event: (1) pre-extreme event, (2) during extreme event, (3) maintenance period, and (4) normal operation.

ESS covers the essential load (50% of the total load). After the recovery time, the MG resumes its normal operation.

V. Conclusion and Future Work

This paper formulates the power management of a MG as a MDP, and solves the sequential decision-making problem using two algorithms: VI and QL. The three agents, controlling the DG, CLs, and the ESS have been trained to find the solution to the proposed multiobjective problem, with the primary focus on maximizing the load coverage and a secondary focus on minimizing the operation cost for a day-ahead finite-horizon.

In case of loss of generation due to an extreme event, the agents would cooperate to make the MG survive with partial load coverage, thus enhancing the resiliency of the system.

In our future studies, we will investigate the memory needed for real time application of the proposed techniques, and the power management of multiple MGs.

References

[1] A. Chaouachi, R. Kamel, R. Andoulsi, and K. Nagasaka, "Multiobjective intelligent energy management for a microgrid," *IEEE Trans. Indust. Elec.*, vol. 60, pp. 1688–1699, Apr. 2013.

[2] R. Lasseter, "Microgrids," *Proc. IEEE Power Eng. Society Trans. and Dist. Conf.*, vol. 1, pp. 305 – 308, Jan. 2002.

[3] C. Schwaegerl, "Advanced architectures and control concepts for more microgrids," *More Microgrids, Siemens AG, STREP*, no. 1, pp. 1–145, 2009.

[4] A. Kaur, J. Kaushal, and P. Basak, "A review on microgrid central controller," *Renew. Sustain. Energy Rev.*, vol. 55, pp. 338–345, 2016.

[5] W. Su, J. Wang, and J. Roh, "Stochastic energy scheduling in microgrids with intermittent renewable energy resources," *IEEE Trans. Smart Grid*, vol. 5, pp. 13–27, 2013.

[6] T. Nguyen and M. Crow, "Stochastic optimization of renewable-based microgrid operation incorporating battery operating cost," *IEEE Trans. Power Syst.*, vol. 31, pp. 2289 – 2296, 2015.

[7] Y. Zhang, N. Gatsis, and G. Giannakis, "Robust energy management for microgrids with high-penetration renewables," *IEEE Trans. Sustain. Energy*, vol. 4, 2012.

[8] D. Jiang, T. Pham, W. Powell, D. Salas, and W. Scott, "A comparison of approximate dynamic programming techniques on benchmark energy storage problems: Does anything work?" *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pp. 1–8, 2014.

[9] P. Zeng, H. Li, H. He, and S. Li, "Dynamic energy management of a microgrid using approximate dynamic programming and deep recurrent neural network learning," *IEEE Trans. Smart Grid*, vol. PP, pp. 171–188, 2018.

[10] S. Zhou, Z. Hu, W. Gu, M. Jiang, and X.-P. Zhang, "Artificial intelligence based smart energy community management: A reinforcement learning approach," *CSEE J. Power Energy Syst.*, vol. 5, 03 2019.

[11] M. N. Alam, S. Chakrabarti, and X. Liang, "A benchmark test system for networked microgrids," *IEEE Trans. Ind. Informat.*, vol. 16, no. 10, pp. 6217–6230, 2020.

[12] A. G. Barto, S. J. Bradtke, and S. P. Singh, "Learning to act using real-time dynamic programming," *Artif. Intell.*, vol. 72, no. 1–2, 1995.

[13] C. Essayeh, M. Raiss El-Fenni, and H. Dahmouni, "Towards an intelligent home energy management system for smart microgrid applications," *2016 International Wireless Communications and Mobile Computing Conference (IWCMC)*, pp. 1051–1056, 2016.

[14] S. A. Pourmousavi Kani, M. H. Nehrir, and R. Sharma, "Multi-timescale power management for islanded microgrids including storage and demand response," *IEEE Trans. Smart Grid*, vol. 6, pp. 1185–1195, 2015.

[15] K. Dehghanpour and H. Nehrir, "A market-based resilient power management technique for distribution systems with multiple microgrids using a multi-agent system approach," *Electric Power Comp. Syst.*, vol. 46, no. 16-17, pp. 1744–1755, 2018.

[16] B. Moran, "Microgrid load management and control strategies," *IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*, pp. 1–4, 2016.

[17] K. Dehghanpour, M. H. Nehrir, J. W. Sheppard, and N. C. Kelly, "Agent-based modeling of retail electrical energy markets with demand response," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3465–3475, 2018.

[18] R. Bellman, "A markovian decision process," *Journal of Mathematics and Mechanics*, vol. 6, no. 5, pp. 679–684, 1957.

[19] E. Alpaydin, *Introduction to Machine Learning*, 3rd ed. MIT Press, 2014.

[20] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. dissertation, King's College, Cambridge, UK, 1989.

[21] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.