Branching Dueling Q-Network-Based Online Scheduling of a Microgrid With Distributed Energy Storage Systems

Hang Shuai, *Member, IEEE*, Fangxing Li, *Fellow, IEEE*, Héctor Pulgar-Painemal, *Senior Member, IEEE*, and Yaosuo Xue, *Senior Member, IEEE*

***Abstract*—This letter investigates a Branching Dueling Q-Network (BDQ) based online operation strategy for a microgrid with distributed battery energy storage systems (BESSs) operating under uncertainties. The developed deep reinforcement learning (DRL) based microgrid online optimization strategy can achieve a linear increase in the number of neural network outputs with the number of distributed BESSs, which overcomes the curse of dimensionality caused by the charge and discharge decisions of multiple BESSs. Numerical simulations validate the effectiveness of the proposed method.**

***Index Terms*—Deep reinforcement learning (DRL), distributed energy storage, microgrid optimization, uncertainty.**

I. Introduction

MICROGRID is a promising concept for addressing the challenges of integrating distributed renewable energy and energy storage systems into power networks. Online optimization, which schedules the operation of microgrids according to the real-time state of the system, is a key technique to ensure the economic operation of microgrids.

However, the uncertainties of renewable energy bring great challenges to the online optimization of microgrids. To address this problem, researchers have proposed several online optimization methods, such as model predictive control (MPC) [1], and approximate dynamic programming (ADP) based algorithm [2]. Nevertheless, the online optimization performance of the above methods relies on forecasting information. So, the performance is affected by the forecasting accuracy of renewable energy and load power. To decrease the dependence on forecasting, several other online optimization approaches for microgrids have been proposed, including the Lyapunov optimization [3], the CHASE algorithm [4], and the recently developed deep reinforcement learning (DRL) based optimization methods (e.g., deep Q Network (DQN) [5], MuZero [6]).

Manuscript received February 6, 2021; revised June 11, 2021; accepted July 21, 2021. Date of publication August 9, 2021; date of current version October 21, 2021. This work was supported in part by the U.S. National Science Foundation (NSF) ECCS Awards under Grant 1809458 and Grant 2033910; in part by the CURENT which is an Engineering Research Center funded by NSF and U.S. Department of Energy (DOE) through NSF Award under Grant EEC-1041877; and in part by the U.S. DOE, Office of Energy Efficiency and Renewable Energy and Office of Electricity under Contract DE-AC05-00OR22725. Paper no. PESL-00029-2021. *(Corresponding author: Fangxing Li.)*

Hang Shuai, Fangxing Li, and Héctor Pulgar-Painemal are with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996 USA (e-mail: fli6@utk.edu).

Yaosuo Xue is with the Electrification and Energy Infrastructures Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSG.2021.3103405.

Digital Object Identifier 10.1109/TSG.2021.3103405

Compared with the conventional microgrid online optimization approaches (e.g., MPC), DRL based algorithms learn to operate the system via historical renewable power generation and load sequences, and can make near-optimal scheduling without using any forecasting information [6]. However, the above-mentioned works mainly focus on the online optimization of a microgrid with a single battery energy storage system (BESS), which fails to address the distributed location characteristic of BESSs. With the rapid development of commercial and home energy storage techniques, plenty of BESSs will be installed in distributed locations of microgrids. The huge action space introduced by multiple BESSs brings great challenges to the discrete-action based DRL optimization methods. For instance, the number of actions that need to be explicitly represented in the DQN [5] or MuZero [6] based agents grows exponentially with an increasing number of BESSs. As a result, the DRL based optimization approaches proposed in [5], [6] are difficult to adapt for a microgrid with distributed BESSs.

To overcome the drawbacks of discrete-action based DRL optimization methods for microgrids mentioned above, this letter develops a novel Branching Dueling Q-Network (BDQ) [7] based online optimization strategy for a microgrid with distributed BESSs, which is the main contribution of this work. The designed BDQ based intelligent agent contains a shared decision module followed by several network branches, one for each BESS. The advantage of the developed algorithm is that it can achieve a linear increase of the number of neural network outputs with the number of distributed BESSs, which will provide great scalability and increase the applicability of the algorithm. In addition, to accommodate the characteristics of historical renewable energy power generation and load power sequences, a long short-term memory (LSTM) based shared decision module architecture is designed for the BDQ agent in this letter to extract features from historical data.

This letter is organized as follows. Section II formulates the microgrid online optimization problem as a mixed integer second-order cone programming (MISOCP) problem by adopting a branch power flow model. In Section III, the BDQ based online optimization algorithm for the microgrid is designed. The numerical simulations are presented in Section IV. Section V concludes this work.

II. Optimization Model of the Microgrid

The microgrid investigated in this letter works in a grid connection mode and consists of electric loads, BESSs, controllable distributed generators (DGs) (e.g., diesel generators), and uncontrollable DGs (e.g., PV panel systems and wind turbines). The goal of online optimization is to minimize the operation cost of the microgrid over the optimization horizon under the necessary constraints. The objective function consists of the fuel cost of controllable DGs, the power

exchange cost of the microgrid and the utility grid, the degradation cost of BESSs, and the renewable energy curtailment cost. The operational constraints considered in this work include the power generation limit and ramp rate constraints, the power exchange limit between the microgrid and the utility, the charge/discharge power limit of BESSs, the branch power flow constraints, etc. The details of the microgrid optimization model can be found in [6, eq. (1)–(25)].

III. BDQ Based Online Optimization Strategy for a Microgrid With Distributed BESSs

The decision variables of the online optimization problem include the complex power generation of controllable DGs, PV panels, and wind turbines; the charge and discharge power of distributed BESSs; the complex power exchange between the microgrid and the utility grid; the branch current; the bus voltage; etc. However, the high-dimensional continuous actions force us to face the curse of dimensionality when applying the reinforcement learning methods to solving our problem. To this end, we develop the BDQ [7] based online optimization approach for a microgrid with distributed BESSs, as illustrated in Fig. 1. The BDQ agent only determines the charge and discharge power of distributed BESSs, while the remaining decisions are obtained by solving the single-time period optimal power flow (OPF) subproblem. The advantage of the proposed optimization architecture is that it can operate the system without dependence on any renewable and load power prediction information.

The designed network architecture of the BDQ agent is also given in Fig. 1. The shared decision module consists of three LSTM units and a fully connected network. The LSTM units extract features from load power and renewable energy power sequences, then the extracted features concatenate with the current state of the microgrid and are then fed into a multilayer network. The features computed by the shared decision module are then used to compute the state value and the state-dependent action advantages on the subsequent independent branches [7]. Note that each branch corresponds to a BESS in this work. The state value and the state-dependent action advantages are combined and input to neural networks to compute the Q-values for each BESS charge and discharge dimension. We discretize the charge and discharge decision of each BESS into n feasible values. The individual branch's Q-value at state s when taking decision P_d^b can be given by:

$$Q_d\left(s, P_d^b\right) = V(s) + \left(A_d\left(s, P_d^b\right) - \frac{1}{n} \sum_{P_d^{b'} \in \mathcal{X}_d} A_d\left(s, P_d^{b'}\right) \right) \tag{1}$$

where $d \in \{1, 2, \ldots, N\}$ represents the dth BESS; $V(s)$ is the state value output by the shared decision module; $A_d(s, P_d^b)$ is the state-dependent action advantage, and $P_d^b \in \mathcal{X}_d$. $\mathcal{X}_d$ represents the feasible action space of the dth BESS.

The neural network weights of the BDQ agent are updated by minimizing the following loss function:

$$L = \Xi_{\left(s, P^b, r, s'\right) \sim D} \left[\frac{1}{N} \sum_d \left(y_d - Q_d\left(s, P_d^b\right) \right)^2 \right] \tag{2}$$

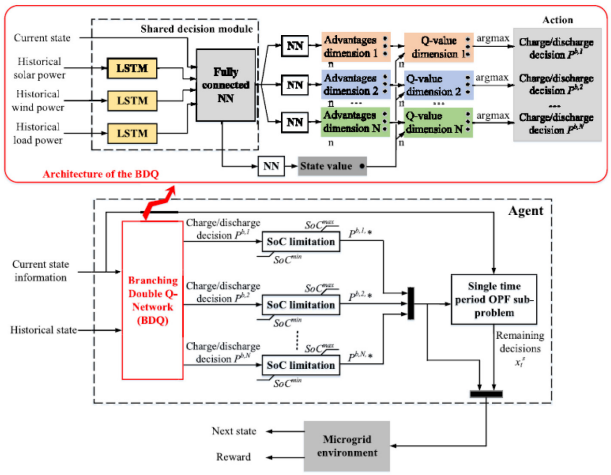

Fig. 1. The developed BDQ based online optimization strategy for a microgrid with distributed BESSs.

where y_d is the temporal-difference (TD) target for the BDQ agent which can be computed by:

$$y_d = r + \gamma \frac{1}{N} \sum_d Q_d^- \left(s', \arg \max_{P_d^{b'} \in \mathcal{X}_d} Q_d\left(s', P_d^{b'}\right) \right) \tag{3}$$

where r represents the reward after taking decision P^b; γ is the discount factor. The details of the training process of the developed BDQ based online optimization algorithm for a microgrid with multiple BESSs is shown in Algorithm 1.

IV. Case Study

To demonstrate the effectiveness of the proposed BDQ based microgrid optimization algorithm, we tested the performance of the algorithm on a 6-bus microgrid, a modified IEEE 33-bus microgrid, and a modified IEEE 69-bus microgrid. All the simulations are conducted on an Intel Core i7-8650U @1.90GHz Windows based computer with 16GB RAM.

The topology of the adopted 6-bus microgrid can be found in [6], and the utilized training and testing dataset of solar power, wind power, load power, and electricity price are the same as in [6]. Although the 6-bus microgrid only contains one BESS, the proposed BDQ based optimization algorithm is also suitable. The convergence process of the proposed algorithm is shown in Fig. 2. From the result, the total returns optimized by the BDQ algorithm approaches the optimal value optimized by the MISOCP method under the condition of perfect information. Note that the MISOCP method needs to know the accurate renewable generation and load power information of all the future time steps, so the optimal objective can be achieved by the method.

To test the online optimization performance of the proposed algorithm, we compared the BDQ based algorithm with several state-of-the-art online optimization algorithms. Using the results optimized by myopic policy as the baseline, the performance improvement of the methods is shown in Table I. We find that the proposed online optimization algorithm outperforms the Lyapunov optimization, ADP, and Deep Deterministic Policy Gradient (DDPG) based optimization algorithms. Although the proposed algorithm performs worse than the MuZero based online optimization method proposed

Algorithm 1 The Training Process of the BDQ Based Online Optimization Algorithm for Microgrid

Initialize the neural networks of the BDQ agent; Initialize experience replay memory; Set the total number of episode N_e and the training frequency f_n, and set the training step $n_{step} = 0$.
for $n_e \leq N_e$ **do**
  Randomly select a day of renewable energy and load sequences from the training data.
  for $t = \Delta t, 2\Delta t, \cdots, T$ **do**
    Get the current state information of the microgrid s_t, and the previous H hours of solar, wind, and load power.
    Compute the charge/discharge decisions of the BESSs using the BDQ agent.
    Recompute the charge/discharge decisions using ϵ-greedy policy.
    Check overcharge/overdischarge limits and get the optimal decisions $P_d^{b,*}(t)$ $(d \in \{1, 2, \cdots, N\})$.
    Solve the OPF sub-problem to get the remaining decisions.
    Execute the optimal decisions x_t to obtain the reward r_t, and calculate the next state of the system $s_{t+\Delta t}$.
    Store the data $(s_t, x_t, r_t, s_{t+\Delta t})$ in the replay buffer.
    if $(n_{step} \ \% \ f_n = 0)$ **then**
      Sample a minibatch of data from the replay buffer.
      Update the main network weights of the BDQ agent to minimize the loss function.
      Update priorities of sampled data.
    $n_{step} = n_{step} + 1$.
    Update target network periodically.
  if $(n_e \ \% \ 500 = 0)$ **then** ▷ *Evaluate every 500 episodes.*
    Evaluate the optimization performance of the BDQ agent.
  $n_e = n_e + 1$.
Return the well-trained BDQ agent parameters.

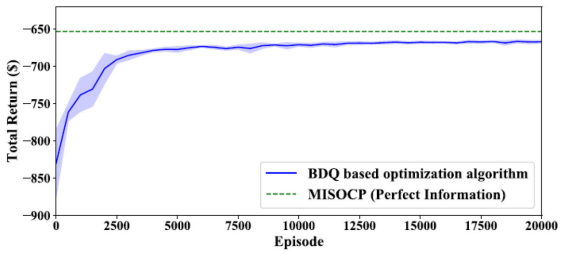

Fig. 2. The convergence process of the proposed BDQ based online optimization algorithm on the 6-bus microgrid system. Blue solid line indicates median returns across 5 separate training runs. The yaxis represents the average total returns for the 10 validation days.

in [6], the MuZero based algorithm is difficult to apply in microgrids with multiple BESSs since the tree search space increases exponentially with the number of BESSs.

To validate the proposed algorithm's ability to solve the online scheduling of microgrids with multiple BESSs, the modified IEEE 33-bus microgrid system was designed as shown in Fig. 3, which contains 5 distributed BESSs. Similarly, the online optimization performance of the proposed algorithm is evaluated and compared with the state-of-the-art methods. The results are given in Table II. Note that the MuZero based approach and look-up table ADP method face the curse of dimensionality due to the huge action space brought by multiple BESSs. Thus, the comparable methods include only the Lyapunov optimization and the DDPG method. Besides, the average time consumption of the BDQ based algorithm, Lyapunov optimization, DDPG, and myopic

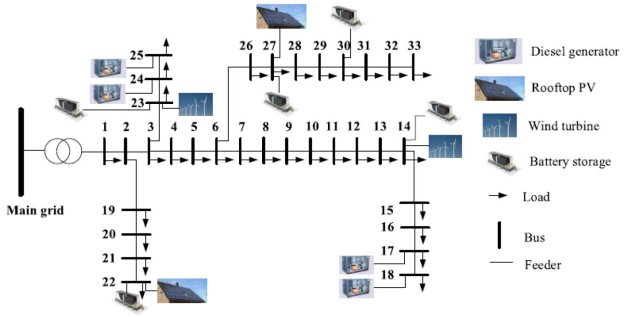

Fig. 3. The diagram of the modified IEEE 33-bus microgrid system.

TABLE I
PERFORMANCE OF DIFFERENT METHODS COMPARED TO MYOPIC POLICY ON THE 100-DAY TESTING DATASET FOR THE 6-BUS MICROGRID

Performance improvement		*Mean*	*Maximum*	*Minimum*	*Standard deviation*
Online methods	BDQ based optimization	8.52%	19.94%	3.19%	2.65%
	MuZero based optimization	9.30%	16.68%	5.28%	2.12%
	Lyapunov optimization	3.76%	9.89%	1.93%	1.65%
	ADP	6.57%	14.78%	4.16%	1.92%
	DDPG	5.55%	9.81%	-8.41%	4.16%
Off-line method	MISOCP (perfect information)	10.20%	23.28%	6.45%	3.02%

TABLE II
PERFORMANCE OF DIFFERENT METHODS COMPARED TO MYOPIC POLICY ON THE 100-DAY TESTING DATASET FOR THE IEEE 33-BUS MICROGRID

Performance improvement		*Mean*	*Maximum*	*Minimum*	*Standard deviation*
Online methods	BDQ based optimization	7.48%	13.48%	4.26%	1.97%
	Lyapunov optimization	3.51%	6.88%	2.22%	1.08%
	DDPG	6.64%	15.27%	2.56%	2.82%
Off-line method	MISOCP (perfect information)	10.20%	23.40%	6.58%	2.98%

policy to make one single time-step scheduling are 0.0438s, 0.782s, 0.035s, and 0.676s, respectively. It can be found that the proposed algorithm performs better than the compared online optimization methods, and the scheduling results of the proposed algorithm are near the optimal value optimized by the MISOCP method under the condition of perfect information.

To validate the scalability of the proposed BDQ based scheduling algorithm, the simulations on a modified IEEE 69-bus microgrid system with different number of distributed BESSs were conducted. The topology and parameters of the microgrid system can be found in the 'case69.m' file of MATPOWER. In the modified system, there are six diesel generators which are connected to buses 17, 18, 24, 30, 40, and 58, respectively. Three distributed PV systems are connected to buses 22, 27, and 45, respectively. Three wind turbines are connected to buses 34, 50, and 59, respectively. The parameters of DGs and the BESS can be found in [6]. The performance of the BDQ based scheduling algorithm was tested when there are 1, 3, 5, and 7 BESSs in the microgrid system. The simulation results are shown in Fig. 4. It can be found that the scheduling results of the BDQ based optimization strategy are near the optimal values optimized by MISOCP method under perfect information. And the time

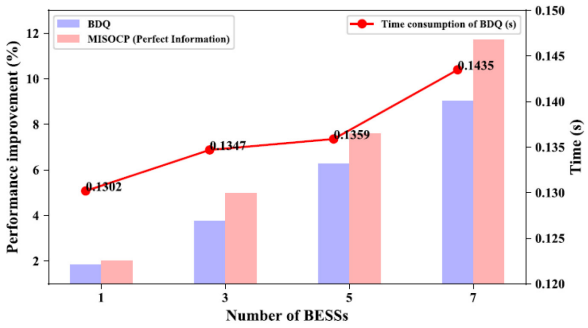

Fig. 4. Performance improvements of BDQ algorithm and MISOCP method compared to myopic policy on the modified IEEE 69-bus microgrid.

consumption of the BDQ agent to make one single time-step scheduling increases linearly with the number of BESSs. Note that the performance improvement of the BDQ based scheduling strategy and MISOCP method increase with the number of BESSs. This can be attributed to the increase in the market arbitrage capacity and renewable energy integration capacity of the microgrid as the energy storage capacity increases, so the gap between the solution of myopic policy and the optimal solution (under perfect information) becomes larger. In addition, we also compared the performance improvement of different methods on the modified IEEE 69-bus microgrid system with 5 BESSs. The average performance improvement of the proposed method, Lyapunov optimization, DDPG, and MISOCP are 6.28%, 2.56%, 5.45%, and 7.61%, respectively.

From the above simulations, the effectiveness and scalability of the proposed BDQ based online scheduling algorithm were validated. Specifically, the proposed BDQ based microgrid scheduling algorithm outperforms many state-of-the-art online scheduling strategies for microgrids in terms of optimization performance and time-consumption.

V. Conclusion

A novel BDQ based online optimization algorithm for microgrids with multiple BESSs was proposed in this letter. The proposed approach enables the linear growth of the total number of agent outputs with increasing BESSs, which provides great scalability and increases the applicability of the algorithm. The simulations indicate that the online optimization performance of the proposed BDQ based approach outperforms the state-of-the-art online optimization methods, such as Lyapunov optimization, ADP, DDPG based method, and MuZero based method. The easy implementation of the algorithm gives it a good application prospect.

References

[1] W. Gu, Z. Wang, Z. Wu, Z. Luo, Y. Tang, and J. Wang, "An online optimal dispatch schedule for CCHP microgrids based on model predictive control," *IEEE Trans. Smart Grid*, vol. 8, no. 5, pp. 2332–2342, Sep. 2017.

[2] H. Shuai, J. Fang, X. Ai, J. Wen, and H. He, "Optimal real-time operation strategy for microgrid: An ADP-based stochastic nonlinear optimization approach," *IEEE Trans. Sustain. Energy*, vol. 10, no. 2, pp. 931–942, Apr. 2019.

[3] W. Shi, N. Li, C.-C. Chu, and R. Gadh, "Real-time energy management in microgrids," *IEEE Trans. Smart Grid*, vol. 8, no. 1, pp. 228–238, Jan. 2017.

[4] Y. Jia, X. Lyu, P. Xie, Z. Xu, and M. Chen, "A novel retrospect-inspired regime for microgrid real-time energy scheduling with heterogeneous sources," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 4614–4625, Nov. 2020.

[5] V. François-Lavet, D. Taralla, D. Ernst, and R. Fonteneau, "Deep reinforcement learning solutions for energy microgrids management," in *Proc. Eur. Workshop Reinforcement Learn. (EWRL)*, 2016, pp. 1–7.

[6] H. Shuai and H. He, "Online scheduling of a residential microgrid via Monte-Carlo tree search and a learned model," *IEEE Trans. Smart Grid*, vol. 12, no. 2, pp. 1073–1087, Mar. 2021.

[7] A. Tavakoli, F. Pardo, and P. Kormushev, "Action branching architectures for deep reinforcement learning," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 4131–4138.