Proceedings of the 27th International Conference on Automation & Computing, University of the West of England, Bristol, UK, 1-3 September 2022

Advanced Machine Learning Approach of Power Flow Optimization in Community Microgrid

Jamal Aldahmashi[1,2], Xiandong Ma[1]
[1]Engineering Department, Lancaster University, Lancaster LA1 4YW, UK
[2]Department of Electrical Engineering, College of Engineering, Northern Border University, Arar 91431, Saudi Arabia
j.aldahmashi@lancaster.ac.uk, xiandong.ma@lancaster.ac.uk

***Abstract*— With the increasing penetration of distributed renewable energy (DERs), the electrical grid is experiencing, on a daily basis, rapid and massive fluctuations in power and voltage profiles. Fast and precise control strategies in real-time have played an important role to ensure that the power system operates at an optimal status. Solving real-time optimal power flow (OPF) problems while satisfying the operational constraints of the community microgrid (CMG) is considered a promising technique to control the fluctuations of renewable sources and loads. This paper adopts a new deep reinforcement learning algorithm (DRL), called Twin-Delayed Deep Deterministic Policy Gradient (TD3), to solve the real-time OPF with consideration of DERs and distributed energy storages (DESs) in the CMG. Training and testing of the algorithm are conducted on an IEEE 14-bus test system. Comparative results show the effectiveness of the proposed algorithm.**

Keywords: Reinforcement Learning (RL), Deep Reinforcement Learning (DRL), Optimal Power Flow (OPF), Twin-Delayed Deep Deterministic Policy Gradient (TD3), Community Microgrid (CMG).

I. INTRODUCTION

Community microgrids (CMGs) have emerged as an effective method to accommodate new energy sources like solar photovoltaics (PVs) and wind turbines (WTs) in the power system. The high penetration of these sources into the power system has become clearly observed due to climate change, and environmental protection. Distributed energy resources (DERs), like PVs and WTs, are distinguished by their uncontrollability and intermittency, thus making the operator of the grid not be able to control and predict their generation. The penetration of DERs has formed huge challenges to the grid operators to ensure that the modern power systems operate at high efficiency and reliability. Dealing with these challenges requires a faster solution to the high-dimension non-convex real-time OPF problems.

In recent years, a number of promising methods have been proposed to solve the OPF problem. Reference [1] uses the swarm intelligent method to solve the OPF problem. However, the nonlinear characteristic of the loads, generators, and other devices that connect to the system makes this optimization method inefficient to find the optimum solution for the OPF problem [2]. On the other hand, evolutionary optimization algorithms have emerged as techniques to transact with the nonlinear characteristics of the OPF problem. However, these techniques require plenty of time or a relatively small space of policies to reach the optimal solution [3].

Reference [4] proposes a stochastic optimization (SO) method to solve the OPF problem. This method depends on finding the distribution of uncertain variables like loads and PVs in the OPF problem. However, the SO requires a large amount of time to calculate a number of scenarios that are used to form the uncertainty distribution [5]. In contrast to the SO, model predictive control (MPC) algorithm is proposed in [6] to solve the multi-phase OPF problem. The MPC algorithm is a very efficient method to reach an optimum solution; however, the performance of the algorithm relies mainly on the prediction precision of the load and renewable energy generation in the power grids, which is difficult to achieve in reality [7].

On the other side, machine learning (ML) has gained wide popularity in recent years, due to its superior ability to take fast decision making even with existing high uncertain variables in the power systems. ML is able to extract useful information from historical data and uses this information to solve an OPF problem. Among the ML approaches, reinforcement learning (RL) is considered an efficient approach to tackle the real-time OPF problem due to its capability to learn the best strategies to reach optimum solutions by searching historical data [8].

Reference [9] proposes Double Deep Q Learning (DDQN), which is a deep RL (DRL) algorithm to solve the stochastic OPF by considering the uncertainty of the load demand and wind turbines. The deep neural network (DNN) that is used in this reference works as an action-value function approximator. However, the DDQN works by discretizing the action space. Since the OPF is a continuous optimization problem, the discretization of the action space leads to loss of some information; thus, DDQN may converge to suboptimal solutions [10].

A number of RL algorithms have been used to deal with continuous environments. In [11], Deep Deterministic Policy Gradient (DDPG) is used to solve real-time OPF by adjusting the active and reactive power of the generators. The proposed approach is tested on the IEEE 118-bus system and the simulation results show that the DDPG reached the optimum solution much faster than the interior-point method. Reference [12] uses a state-of-art ML method which is based on Deep Policy Gradient (DPG), called Proximal Policy Optimization (PPO), to solve real-time OPF. The PPO is highly effective to work with high-dimensional and continuous action space like OPF. Moreover, PPO is able to take multiple actions at the same time, which gives this method superiority over the traditional ML methods. The proposed approach is tested on the IEEE 33-bus system, and the simulation results

demonstrate that PPO can provide a more resilient control strategy than the stochastic programming method.

Inspired by the recent researches, this paper presents a new ML algorithm with continuous action search to solve real-time alternating current (AC) OPF (ACOPF) in the CMG that is equipped with DERs and distributed energy storages (DESs). The real-time ACOPF problem is formulated as Markov decision process (MDP), and then the twin-Delayed Deep Deterministic Policy Gradient (TD3) is used to solve the MDP. This paper takes into consideration the uncertainty of the WTs, PVs, and load demand, and the objective is to minimize the total power loss by controlling the active and reactive power of the DERs and DESs. The RL agent has been developed to ensure the CMG is operating under safe and reliable conditions, meaning that the agent actions must satisfy all operational constraints.

II. PROBLEM FORMULATION

A CMG can be represented as a set of buses $\mathcal{N}$ that are connected with each other by transmission lines (TLs), where $\mathcal{N}$ is a set of positive integers. The TLs can be represented as a matrix $\mathcal{E} = \mathcal{N} \times \mathcal{N}$ that demonstrates the relation between buses within a CMG. All CMG devices $\mathcal{D}$ are connected to single or several buses, which might withdraw power from the grid, e.g., loads, or inject power into the grid, e.g., DERs. Each bus $i \in \mathcal{N}$ in the CMG can be represented by several variables that determine its status, including a current I_i, voltage level V_i, active power $P_i^{(\text{bus})}$ and reactive power $Q_i^{(\text{bus})}$. The active or reactive power injection into or withdrawal from the bus can be obtained by calculating the active and reactive power of the device $\mathrm{d} \in \mathcal{D}$ that connects with it. The devices in the CMG are divided into three subsets $\mathcal{D}_{DER}$, $\mathcal{D}_L$,and $\mathcal{D}_{DES}$. $\mathcal{D}_{DER}$ represent the DER that can only inject power into the CMG, while $\mathcal{D}_L$ is passive loads that can only withdraw power from the CMG. On the other hand, $\mathcal{D}_{DES}$ is a storage unit that can inject and consume power into/from the grid. The proposed CMG works in connection with the main grid, which is used to provide a voltage reference and also to balance the power level in the CMG.

The objective function of the OPF problem, as shown in (1), is to minimize the power loss while the security constraints are satisfied. The optimization horizon is one day, and the time interval (timestep) is a 15- minutes, and thus 96 time steps a day.

$$\min \sum_{t=0}^{96-1} P_{\mathcal{D}_L,loss} + P_{\mathcal{D}_{DER},loss} + P_{\mathcal{D}_{DES},loss} \quad (1)$$

where $P_{\mathcal{D}_L,loss}$ is the total transmission energy loss that occurs as a result of leakage in TLs. $P_{\mathcal{D}_{DES},loss}$ is the total energy loss that happens due to leakage in storage units. This loss occurs because DESs have charging and discharging efficiency factors η. $P_{\mathcal{D}_{DER},loss}$ is the total energy loss that happens when the generation of DER is curtailed. This curtailment is necessary when the generation power of the DER is higher than the required power capacity of TLs. $P_{\mathcal{D}_L,loss}$, $P_{\mathcal{D}_{DES},loss}$ and $P_{\mathcal{D}_{DER},loss}$ can be calculated by using (2), (3), and (4), respectively.

$$P_{\mathcal{D}_L,loss} = \sum_{d \in \mathcal{D}_L} P_{d,t+1} \quad (2)$$

$$P_{\mathcal{D}_{DES},loss} = \sum_{d \in \mathcal{D}_{DES}} P_{d,t+1}\,(1-\eta) \quad (3)$$

$$P_{\mathcal{D}_{DER},loss} = \sum_{g \in \mathcal{D}_{DER}} P_{g,t+1}^{(max)} - P_{g,t+1} \quad (4)$$

where $P_{g,t+1}^{(max)}$ is the maximum generation power of the DER, and $P_{g,t+1}$ is the renewable energy production after curtailment. $P_{d,t+1}$ in (2) is transmission energy loss that occurs in TLs as a result of passing power to the loads, while $P_{d,t+1}(1-\eta)$ in (3) is a power loss of the batteries.

The solution of the OPF problem must be considered the DERs, DESs, and network constraints. Equations (5), and (6) represent the physical limitation of DERs and DESs, while (7) and (8) are used to add an additional constraint on the reactive power injection when the active power of the DERs or DESs are close to the maximum value. Equations (9) and (10) represent additional constraints on the active power injection/withdrawal of storage units. The DESs cannot withdraw (charge) an active power when the current state of charge $SoC_{d,t}$ indicates that the storage unit is full $\overline{SoC_d}$. On the other hand, DESs would not be able to inject (discharge) an active power when the $SoC_{d,t}$ is empty $\underline{SoC_d}$.

$$\underline{P_g} \le P_g \le \overline{P}_g \qquad \forall g \in \mathcal{D}_{DER,DES} \quad (5)$$

$$\underline{Q_g} \le Q_g \le \overline{Q}_g \qquad \forall g \in \mathcal{D}_{DER,DES} \quad (6)$$

$$Q_g \le \tau_g^{(1)} P_g + \rho_g^{(1)} \qquad \forall g \in \mathcal{D}_{DER,DES} \quad (7)$$

$$Q_g \ge \tau_g^{(2)} P_g + \rho_g^{(2)} \qquad \forall g \in \mathcal{D}_{DER,DES} \quad (8)$$

$$P_g \ge \frac{1}{\Delta \mathrm{t}\,\eta}\left(SoC_{g,t-1} - \overline{SoC_g}\right) \qquad \forall g \in \mathcal{D}_{DES} \quad (9)$$

$$P_g \le \frac{\eta}{\Delta t}\left(SoC_{g,t-1} - \underline{SoC_g}\right) \qquad \forall g \in \mathcal{D}_{DES} \quad (10)$$

where $\tau_g^{(1)}, \tau_g^{(2)}, \rho_g^{(1)}$, and $\rho_g^{(2)}$ are constant values and Δt is the time difference during $(t, t+1)$

Two types of network constraints are considered in this paper, which must be satisfied all the time. The first constraint is the voltage magnitude of the buses which must be within acceptable limits, as is given in (11). The second constraint, as is shown in (12), represents the maximum value of the apparent power that can flow from bus i to bus j.

$$\underline{V_i} \le V_{i,t} \le \overline{V}_i \qquad \forall i \in \mathcal{N} \quad (11)$$

$$S_{ij,t} \le \overline{S}_{ij,t} \qquad \forall e_{ij} \in \mathcal{E} \quad (12)$$

III. MARKOV DECISION PROCESS OVERVIEW

The MDP is used to model the CMG framework, which can be divided into four parts: $(\xi, \mathcal{A}, \mathbb{R}, P)$, representing state space, action space, reward function and transition function, respectively. The goal of the RL agent is to lean an optimal policy π^* by continuously interacting

with the CMG environment to maximize a cumulative reward as depicted in Fig. 1. As the optimization of the real-time OPF is a sequential decision-making problem, the RL agent receives a vector of state $s_t \in \xi$ and reward $r_{t-1} \in \mathbb{R}$ at timestep t. Based on this information, the agent provides actions $a_t \in \mathcal{A}$ to the CMG environment.

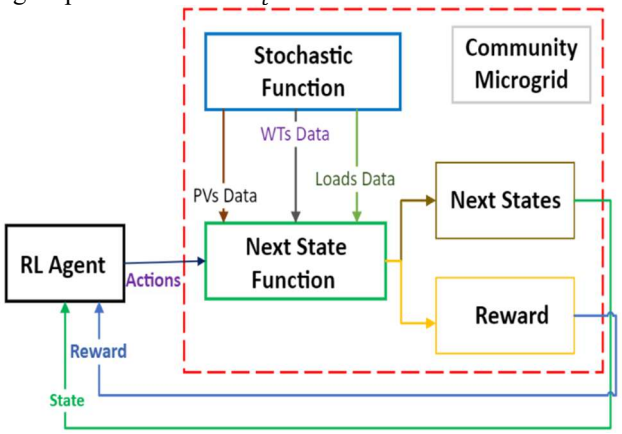

Fig. 1 The environment framework

The environment adds a next load demand $P_{\mathcal{D}_L,t+1}$, and maximum generation of PVs $P^{max}_{PV,t+1}$ and WTs $P^{max}_{WT,t+1}$ before curtailment to the agent actions. Then, the next state function receives a_t, $P_{\mathcal{D}_L,t+1}$, $P^{max}_{PV,t+1}$and $P^{max}_{WT,t+1}$ and uses the Newton-Raphson method to determine all bus voltages of the CMG and the active and reactive power of the main grid.

The next state function updates the values of all the currents, voltages, active and reactive powers of the CMG, resulting in a new state $s_{t+1} \in \xi$ which are sent to the RL agent to process again. Furthermore, it uses (13) to calculate the reward r_t:

$$r_t = \text{clip}\left(\pm r_{\text{clip}}, -\left(\Delta E_{Loss,t:t+1} + \lambda\phi(s_{t+1})\right)\right) \quad (13)$$

where $\Delta E_{Loss,t:t+1}$ is the total power loss, and $\phi(s_{t+1})$ is a penalty term associated with excess limitation of the operating constraints, λ is a weighing hyperparameter. r_{clip} is used to make the reward function within a finite range $\pm r_{\text{clip}}$. The agent uses (13) to learn optimal policy π^* that minimizes the power loss while the operational constraints are satisfied.

A. State space ξ

The state space ξ is used to describe the CMG environment and also used as the input of RL algorithm. $s_t \in \xi$ includes the active and reactive powers $P^{device}_{d,t}, Q^{device}_{d,t}$ of all devices in the environment $d \in \mathcal{D}$, charge level of the storage unit $(SoC_d), d \in \mathcal{D}_{DES}$, the maximum power generated from DERs $P^{(max)}_g$ at time t. The environment reaches a terminal state ξ^{ter} when $t =$ 96 timesteps or Newton-Raphson method could not find a solution to (14) as a result of the actions that are taken by the RL agent.

$$P^{(bus)}_i + i\, Q^{(bus)}_i = V_i I^*_i \qquad \forall i \in \mathcal{N} \quad (14)$$

where I^*_i is the complex conjugate of the current I_i.

B. Action space $\mathcal{A}$

Given the state $s_t \in \xi$ of the CMG at a specific timestep t, actions $a_t \in \mathcal{A}$ are taken by the RL agent to determine the curtailment value of active power or reactive power of the DERs, and to set the active power or reactive power injection into or withdraw from DESs. These actions must satisfy the constraints of DERs, DESs, and network stability all the time.

C. Transition function P

Since the WTs and PVs power generation and load demand for next timestep are stochastic, the state transitions of $\left\{P^{(max)}_{g,t+1}\right\}_{g\in\mathcal{D}_{DER}}$ and $\left\{P_{d,t+1}\right\}_{d\in\mathcal{D}_L}$ are dependent on the CMG randomness. State transition in the CMG occurs in three steps as is shown in Fig.1. Once the agent selects actions a_t, the environment combines the next step power generation and the load with selected actions and passes them to the next state function which maps the next state $s_{t+1} \in \xi$ with current (state s_t, action a_t) pair and determines the reward.

D. Reward function $\mathbb{R}$

$\mathbb{R}$ represents the reward signal after action a_t is taken in s_t and it is defined as:

$$r_t = \begin{cases} \text{clip}\left(\pm r_{\text{clip}}, -\left(\Delta E_{Losst:t+1} + \lambda\phi(s_{t+1})\right)\right) & if\ s_{t+1} \notin \xi^{ter} \\ -\dfrac{r_{\text{clip}}}{1-\gamma} & if\ s_{t+1} \in \xi^{ter} \end{cases} \quad (15)$$

where γ is a constant value.

Using a clipping parameter r_{clip} in reward function is to ensure that any transition from a non-terminal to a terminal state produces a very high negative reward. The objective of the clipping function is to encourage the RL agent to lean a policy π^* that avoids all the scenarios that might make the CMG collapse.

The total power loss $\Delta E_{Loss,t:t+1}$consists of three parts as defined in (2)-(4). Equation (16) is used to represent all the power loss in the CMG.

$$\Delta E_{Loss,t:t+1} = \sum_{d\in\mathcal{D}_L} P_{d,t+1} + \sum_{d\in\mathcal{D}_{DES}} P_{d,t+1}(1-\eta) + \sum_{d\in\mathcal{D}_{DER}} \left(P^{(max)}_{g,t+1} - P_{g,t+1}\right) \quad (16)$$

The penalty term $\phi(s_{t+1})$ adds a high negative reward when the agent violates the operating constraints. The penalty term can be obtained as:

$$\Phi(\mathbf{s}_{t+1}) = \Big(\sum_{i\in\mathcal{N}} \Big(max\left(0, |V_{i,t+1}| - \overline{V}_i\right) + max\left(0, \underline{V}_i - |V_{i,t+1}|\right)\Big) + \sum_{e_{ij}\in\xi} \max\left(0, |S_{ij,t+1}| - \overline{S}_{ij}, |S_{ji,t+1}| - \overline{S}_{ij}\right)\Big) \quad (17)$$

where $|S_{ij,t+1}| \neq |S_{ji,t+1}|$ due to TL loss. The first part of (17) represents the limits of the allowed voltage at each bus, which is necessary to maintain the stability of the CMG. The second part is referred to the maximum rating of the power that can flow in the TLs. This constraint is

essential to prevent TLs from overheating. This reward $\Phi(\mathbf{s}_{t+1})$ is defined to have a higher negative value than energy loss $\Delta E_{Loss,t:t+1}$, because unsatisfying the network constraints can lead to damaging the CMG infrastructure.

IV. OPTIMIZATION METHODS

A. MPC

An MPC algorithm is used to solve the multi-stage OPF problem to evaluate the RL agent's performance in a specific environment. An MPC algorithm can efficiently solve the OPF problem, especially if the prediction of load demand and DERs generation is highly accurate over the optimization horizon.

B. TD3 algorithm

The goal of the TD3 agent is to minimize the total power loss without violating the network constraints or shedding the load. The agent controls the power outputs of DERs and DESs under different load conditions. The RL agent trains through historical data which is called offline training and then applies this model in real-time application.

TD3 is a DRL approach that combines DQN with DDPG. The TD3 is an actor-critic approach that contains both the actor-network and the critic network. The critic network is a Q-value network that takes states and actions as inputs and the output of this network is the estimated Q-value function. The actor-network performs policy improvement to update the policy according to the Q-value function that generates the critic-network. In other words, the actor-network produces an action for the following state. The critic network is in charge to evaluate the policy (policy evaluation) by generating Q-value estimated and computing the difference (temporal difference) between this value and the value generated by the actor-network. Instead of maximizing the Q-value function, the critic network evaluates the gradient of the Q-value to find the orientation of the change action for getting a higher Q-value estimated. Consequently, the actor-network updates its weights in the direction of the gradient of the loss function. TD3 is called "twin" because it uses two critic-networks, two actor networks, and two Bellman equations. Also, it is called "delayed", because the policy is updated less than the Q-value function. This delayed update makes the value estimation have a lower variance. Therefore, lower variance means better policy. TD3 uses action smoothing and utilizes this technique to trade-off between exploitation and exploration. TD3 adds noise to the target action, to make it harder for the policy to exploit Q-function errors by smoothing out Q along with changes in action.

TD3 uses the backpropagation method on the critic loss to determine the parameters of the networks. It updates the parameters of the actor-network every two iterations by implementing the gradient ascent on the output of the first critic network. As is shown in Fig.2, the TD3 agent takes the states of the CMG as inputs and according to optimal policy, the agent provides actions to the environment.

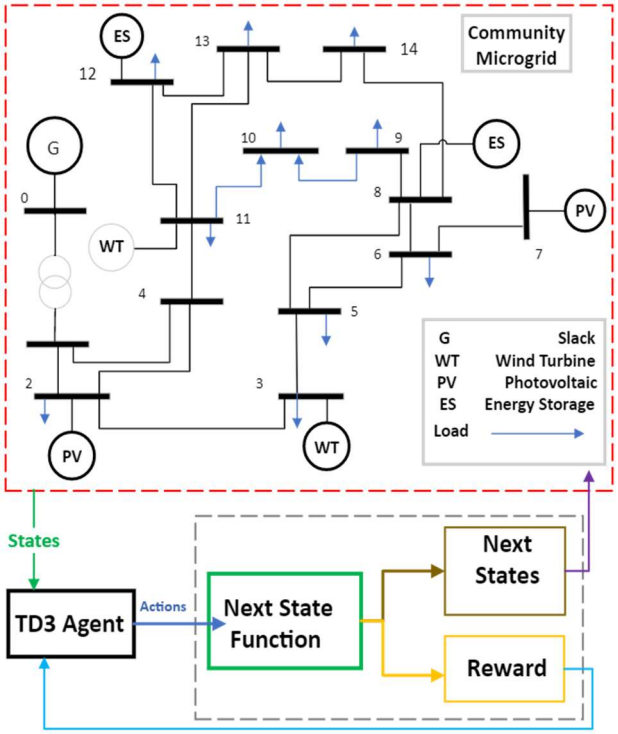

Fig. 2 Interaction process between the TD3 agent and CMG environment.

V. CASE STUDY

A. Network architecture

The proposed approach (TD3) is tested on an open-source OPF environment called Gym-ANM [13], which provides the researchers the ability on designing its own microgrid topology, OPF equations, and modifying the characteristics of loads, DERs, and DESs, and testing RL algorithms. The proposed topology of the CMG used in this paper is shown in Fig.2. The CMG is connected with the main grid, and consists of fourteen buses, where solar PVs and WTs work as DERs in buses 2, 3, 7, and 10, respectively, while, the DESs (ES in Fig. 2) are located in buses 8 and 12. The rated power of the WTs and solar PVs are 50 KW and 35 KW, respectively. The capacity of the installed DESs are 200 KWh. The charging and discharging efficiency η are both set as 90%. The characteristics of all CMG devices is summarized in Table 1.

B. Experimental details

In this section, the performance of TD3 and other DRL algorithms are compared against the MPC algorithm with a perfect forecast of the CMG data. These algorithms are tested on a modified IEEE 14-bus system. Training set is created by generating random data for each load and DER for seven days (672-time steps) as shown in Fig. 3. The RL agents will be trained for 200 episodes with these data. The performance of the RL agents are tested in five random days, which represent the testing set. To make the CMG

model close to reality, the weather is supposed to be cloudy on Day 1, which means the generation of solar PV is zero as is shown in Fig. 4. Moreover, the weather on Day 2 is assumed to be windy which means the wind turbine generate power at the maximum limits as shown in Fig 5. The weather in the rest of the testing set is presumed to be normal as shown in Fig 6. It is worthy to mention that the MPC algorithm is fed by perfect forecast data which means the forecasted data error is 0%. In reality, it is impossible to predict the future load demand and DERs generation with a 0% error rate.

TABLE 1. Descriptions of each device (loads, DERs, DESs)

Number of devices	Type of device	$\overline{P}_d$ (KW)	$\underline{P}_d$ (KW)	$\overline{Q}_d$ (KVAR)	$\underline{Q}_d$ (KVAR)	η
1	Load	-	-20	-	-	-
2	PV	35	0	35	-35	-
3	Load	0	-20	-	-	-
4	WT	50	-50	50	-50	-
5	Load	0	-15	-	-	-
6	PV	35	0	35	-35	-
7	Load	0	-20	-	-	-
8	DES	50	-50	50	-50	0.9
9	Load	0	-20	-	-	-
10	Load	0	-20	-	-	-
11	Load	0	-20	-	-	-
12	WT	50	-50	50	-50	-
13	DES	50	-50	50	-50	0.9
14	Load	0	-20	-	-	-
15	Load	0	-20	-	-	-
16	Load	0	-20	-	-	-

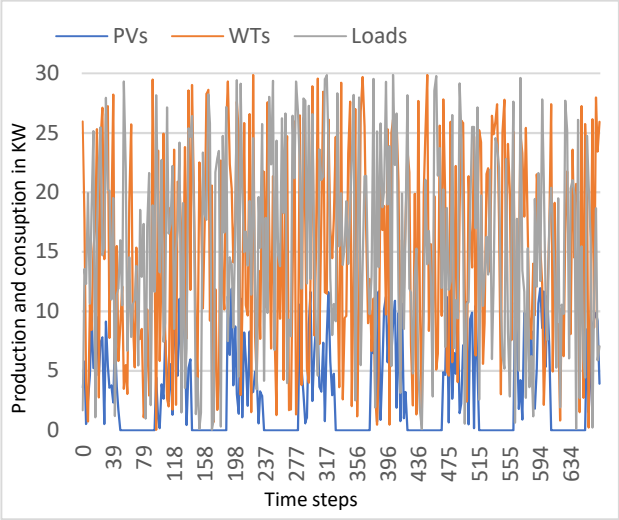

Fig. 3 Cumulative consumption and generation for training data.

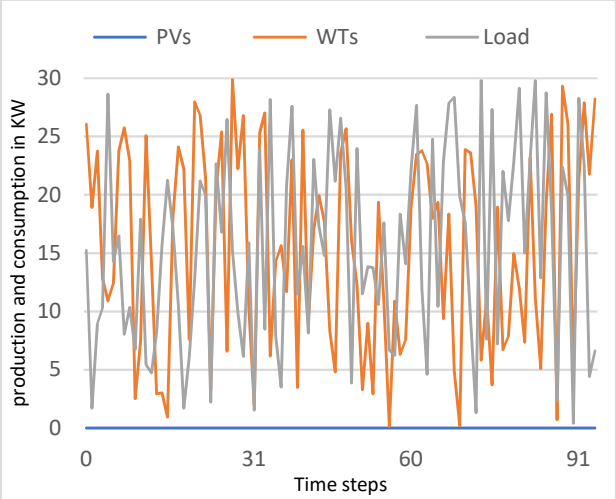

Fig. 4 Cumulative consumption and generation for in Day 1 (PV=0)

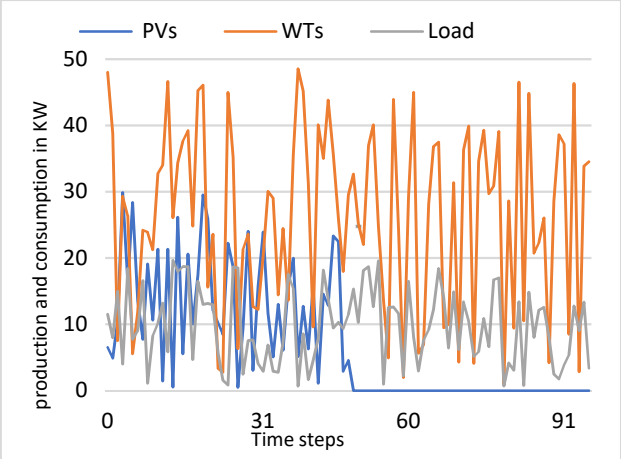

Fig. 5 Cumulative consumption and generation in Day 2 (WT=high).

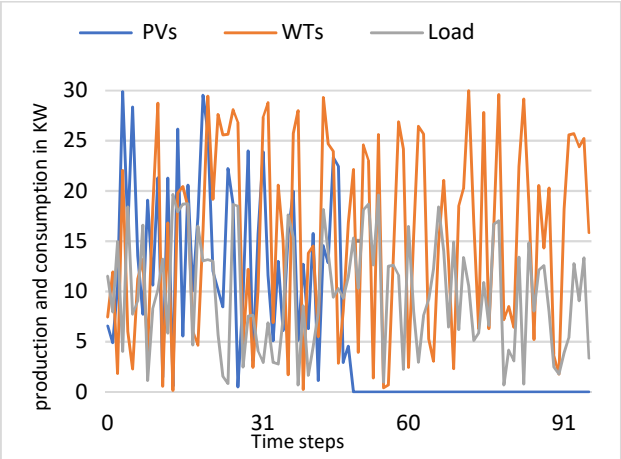

Fig. 6 Cumulative consumption and generation for rest of the testing set.

C. Performance evaluation

Equation (15) is used to evaluate the performance of the proposed algorithms, and we make a comparison with the simulation results between TD3 algorithm and other DRL algorithms, e.g., PPO, soft actor-critic (SAC), DDPG, and Advantage Actor-Critic (A2C). Moreover, we

compare the results of all DRL algorithms against the MPC approach.

D. Comparison results

As shown in Fig 7, among all the DRL algorithms, the proposed approach (TD3) has the minimum power loss, except on the day 3. This indicates that the performance of the TD3 algorithm outperforms A2C, SAC, PPO and DDPG in terms of power loss minimization. MPC results are also given in Fig 7 (black dashed curve), which is plotted as the benchmark to demonstrate the optimal results obtained by the traditional optimization method. Although the performance of the MPC method is better than the TD3 algorithm on all testing days, it is not suitable to solve real-time OPF since it takes much longer time to find an optimal solution. The results show that TD3 is 3.8 times faster than MPC. As highlighted before, MPC cannot work efficiently without having perfect forecasted data.

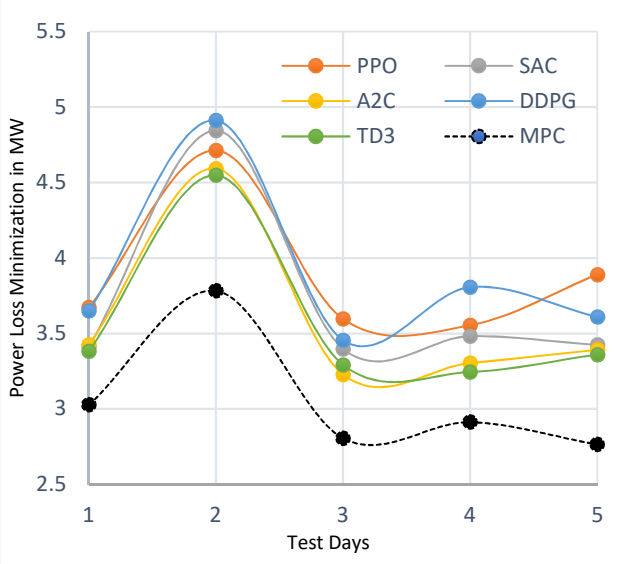

Fig. 7 Power loss minimization results of different approaches for five random days.

VI. CONCLUSION

The increasing penetration of solar PVs, WTs, and energy storages present major challenges for the operation of the CMG. In this paper, DRL based approach is proposed for management of the CMG under uncertainties. The real-time OPF problem is formulated as an MDP, then the TD3 is used to extract optimal operation of the CMG by using DRL from the historical data. The proposed algorithms are tested on IEEE 14-bus system and the simulation results show that the TD3 algorithm outperforms the state-of-art DRL algorithms like PPO and A2C.

Future work includes exploring improvements in the performance of the proposed method in terms of how to enhance the optimality while preserving the feasibility at the same time. We will also use long-time data to train the models and real data for validation of the models. We intend to include hyperparameters optimization and use practical data to test the algorithm in our future work. Furthermore, more realistic system operation conditions should be applied to validate the robustness of the proposed method.

REFRENCES

[1] T. Niknam, H. Z. Meymand, and H. D. Mojarrad, “An efficient algorithm for multi-objective optimal operation management of distribution network considering fuel cell power plants,” Energy, vol. 36, no. 1, pp. 119–132, Jan. 2011.

[2] F. Capitanescu, “Critical review of recent advances and further developments needed in AC optimal power flow,” Electric Power Systems Research, vol. 136, pp. 57–68, Jul. 2016.

[3] R. S. Sutton and A. Barto, Reinforcement learning : an introduction. Cambridge, Ma ; Lodon: The Mit Press, 2018.

[4] Y. Xu, Z. Y. Dong, R. Zhang, and D. J. Hill, “Multi-Timescale Coordinated Voltage/Var Control of High Renewable-Penetrated Distribution Systems,” IEEE Transactions on Power Systems, vol. 32, no. 6, pp. 4398–4408, Nov. 2017.

[5] D. Bertsimas, E. Litvinov, X. A. Sun, J. Zhao, and T. Zheng, “Adaptive Robust Optimization for the Security Constrained Unit Commitment Problem,” IEEE Transactions on Power Systems, vol. 28, no. 1, pp. 52–63, Feb. 2013.

[6] P. Fortenbacher, A. Ulbig, S. Koch, and G. Andersson, “Grid-constrained optimal predictive power dispatch in large multi-level power systems with renewable energy sources, and storage devices,” IEEE Xplore, Oct. 01, 2014.

[7] 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 Transactions on Sustainable Energy, vol. 10, no. 2, pp. 931–942, Apr. 2019.

[8] G. Krishnamoorthy, A. Dubey, and A. H. Gebremedhin, “Reinforcement Learning for Battery Energy Storage Dispatch augmented with Model-based Optimizer,” IEEE Xplore, Oct. 01, 2021.

[9] D. Cao, W. Hu, X. Xu, Q. Huang, and Z. Chen, “Double Deep Q-learning Based Approach for Power Flow Optimization in Distribution Networks,” IEEE Xplore, May 01, 2020.

[10] E. AboElHamd, H. M. Shamma, and M. Saleh, “Dynamic Programming Models for Maximizing Customer Lifetime Value: An Overview,” Advances in Intelligent Systems and Computing, pp. 419–445, Aug. 2019.

[11] Z. Yan and Y. Xu, “Real-Time Optimal Power Flow: A Lagrangian based Deep Reinforcement Learning Approach,” IEEE Transactions on Power Systems, pp. 1–1, 2020.

[12] D. Cao et al., “Deep Reinforcement Learning Based Approach for Optimal Power Flow of Distribution Networks Embedded with Renewable Energy and Storage Devices,” Journal of Modern Power Systems and Clean Energy, vol. 9, no. 5, pp. 1101–1110, 2021.

[13] R. Henry and D. Ernst, “Gym-ANM: Reinforcement learning environments for active network management tasks in electricity distribution systems,” Energy and AI, vol. 5, p. 100092, Sep. 2021.