

ORIGINAL RESEARCH

Open Access



A soft actor-critic deep reinforcement learning method for multi-timescale coordinated operation of microgrids

Chunchao Hu^{1,2,3,5*}, Zexiang Cai¹, Yanxu Zhang^{2,3}, Rudai Yan⁴, Yu Cai^{2,3} and Bowei Cen¹

Abstract

This paper develops a multi-timescale coordinated operation method for microgrids based on modern deep reinforcement learning. Considering the complementary characteristics of different storage devices, the proposed approach achieves multi-timescale coordination of battery and supercapacitor by introducing a hierarchical two-stage dispatch model. The first stage makes an initial decision irrespective of the uncertainties using the hourly predicted data to minimize the operational cost. For the second stage, it aims to generate corrective actions for the first-stage decisions to compensate for real-time renewable generation fluctuations. The first stage is formulated as a non-convex deterministic optimization problem, while the second stage is modeled as a Markov decision process solved by an entropy-regularized deep reinforcement learning method, i.e., the Soft Actor-Critic. The Soft Actor-Critic method can efficiently address the exploration–exploitation dilemma and suppress variations. This improves the robustness of decisions. Simulation results demonstrate that different types of energy storage devices can be used at two stages to achieve the multi-timescale coordinated operation. This proves the effectiveness of the proposed method.

Keywords: Microgrid operation, Hybrid energy storage system, Deep reinforcement learning

1 Introduction

The microgrid is considered a promising self-sufficient energy system to incorporate renewable energy sources (RES) into the main grid. It is defined as a small cluster of distributed generators (DGs) and energy storage systems (ESS). These can operate in grid-connected or isolated modes [1]. Energy management is usually designed for microgrid operation to improve energy efficiency and minimize the operational cost [2]. From the view of microgrid control and operation, energy management can be considered as the high-level (tertiary) control in the hierarchical microgrid control architecture [3]. The objectives of energy management are to decide

the amount of electricity buying (selling) from (to) the main grid and dispatch the available energy resources in a microgrid (e.g., ESS, DGs). In an energy management strategy, the energy storage system plays a key role in compensating for power mismatch. Many kinds of storage devices can be used as an ESS in a microgrid, such as supercapacitors, batteries, and fuel cells. Power density and energy density are the two main metrics for selecting the proper devices. With the increase of RES penetration level in the microgrid, intermittency, uncertainty, and non-dispatchability of RES make it challenging to maintain the supply–demand balance. To address the instantaneous fluctuations of RES and meet different operational requirements, hybrid ESS (HESS), consisting of various types of storage devices (typically battery and supercapacitor), inherits advantages of each type, and therefore can be deployed as an effective infrastructure to support microgrid operation. Different types of energy storage

*Correspondence: huchunchao@139.com

⁵ School of Electrical and Electronic Engineering, South China University of Technology, Guangzhou 510640, China
Full list of author information is available at the end of the article

devices have different characteristics. For instance, a battery usually has a high energy density but low power density, while, in contrast, a supercapacitor has a low energy density but high power density. To fully exploit the HESS, the different types of devices should be optimally coordinated so that the operational cost is reduced.

In previous studies, many algorithms have been proposed for HESS-based microgrid operation. These can be generally categorized into rule-based and optimization-based approaches [4]. Rule-based methods [5–7] are based on experience and empirical evidence, and are widely used for real-time operation because of their simplicity. However, rule-based methods are limited by the short-vision and may not accurately reflect the actual conditions of ESS in the long run. Optimization-based approaches employ advanced optimization theory and can be classified into global (off-line) optimization [8–12] and real-time (on-line) optimization [13–15]. Nevertheless, optimization-based algorithms are complex and usually cause heavy computational burden in online applications [4]. In addition, the uncertainties from RES affect the effectiveness of optimization methods. For example, stochastic optimization suffers from heavy computational burden, and usually requires the probability distributions of RES that are difficult to obtain, while robust optimization has to make the trade-off between economy and conservatism.

Recently, there has been increased interest in using deep reinforcement learning (DRL) to make real-time, sequential, and reliable decisions with the existence of uncertainty. Reference [16] proposes an energy management approach for real-time scheduling of microgrids considering the uncertainty of demand, RES, and electricity price, while [17] establishes a DRL-based energy management strategy in the energy internet. An intelligent multi-microgrid energy management method based on deep learning and model-free reinforcement learning (RL) is introduced in [18]. However, existing works on DRL-based energy management only focus on long or short-term economic dispatch of ESS operation under different conditions without considering the coordination of different types of storage devices. In addition, previous studies mostly adopt the single-stage dispatch model. However, simply implementing a DRL agent in the system may lead to unexpected decisions. In addition, entirely depending on a DRL agent to make decisions lacks reliability in practical operation, and interpretability is usually unsatisfactory.

To address the highlighted problems, this paper proposes the use of a two-stage coordinated approach to guide a DRL agent so that it can select actions within a proper range to improve the reliability of decisions.

A DRL-based multi-timescale coordinated operation method for microgrids with HESS is developed. The microgrid has two types of ESS, i.e., battery with high energy density and supercapacitor with high power density. This method realizes multi-timescale coordination by introducing a hierarchical two-stage dispatch model. The first stage makes an initial decision without considering the RES uncertainties to minimize the operational cost, including the degradation cost and regulation cost from the utility grid. Power references of the battery and electricity from the utility grid are also calculated, and these will be further used in the second stage. The second stage generates corrective actions to compensate for the first-stage decision after the uncertainty is realized. The first stage is a nonlinear programming problem and is solved by an off-the-shelf solver over a longer timescale (1 h), while the second stage is modeled as a Markov decision process addressed by DRL in real-time.

The Soft Actor-Critic (SAC), as an efficient DRL algorithm for a continuous control problem [19], is adopted in the approach. The SAC is a stochastic off-policy actor-critic algorithm based on the concept of entropy, and it returns a probability distribution over the action space. The entropy term is used to evaluate the randomness of the policy, while the SAC introduces an entropy term in the reward function so that the return and exploration can be maximized at the same time during the training process. This improves exploration and robustness. Compared with the deep deterministic policy gradient (DDPG), another effective DRL method [20], the policy learned by the SAC is a stochastic policy that returns a probability distribution over the action space, while DDPG often has larger variations and can easily converge to a local optimum [21]. Thus, the SAC can better address the uncertainty existing in the decision due to its stochastic policy.

The contributions of this paper can be summarized as follows:

- (1) A DRL-based two-stage energy management framework is established for microgrids with HESS. The first stage schedules the output of battery and regulation from the utility grid (buying/selling electricity from/to grid), while the second stage adjusts the scheduling from the first stage and determines the power of the supercapacitor to deal with the instantaneous fluctuations of RES.
- (2) The degradation models of battery and supercapacitor are derived and considered during operation so that the long-term capital cost and short-term operational cost can be coordinated.

- (3) An advanced DRL algorithm, namely the SAC, is derived to improve training efficiency and performance in the second stage. The results show that the SAC agent has low variation and stable performance.

2 Microgrid operation modelling

2.1 Microgrid

Here the microgrid consists of HESS and RES, including photovoltaic (PV) and wind turbine (WT), and aggregated load. The microgrid can interact with the utility grid through the point of common coupling (PCC). Figure 1 shows the schematic of the microgrid for energy management. When the microgrid operates in the grid-connected mode, it can benefit from selling electricity to the utility grid or buying electricity from the grid to maintain the power balance. The power balance to be maintained can be expressed as:

$$P_L = P_{PV} + P_{WT} + P_{Grid} + \underbrace{P_B + P_{SC}}_{P_{ESS}} \quad (1)$$

Among these variables, P_L , P_{PV} and P_{WT} are considered to be uncontrollable, and the objective of energy management is to determine optimal values for P_{Grid} , P_B and P_{SC} so that the microgrid operation cost can be minimized. We also assume that the utility grid adopts a dynamic real-time pricing scheme in which the price is determined based on the bidding of the electricity market participants and is usually available for the public several hours ahead, allowing customers to make a schedule in advance [22]. The price data from the Energy Market Authority (EMA) in Singapore [23] over a year is used in this work.

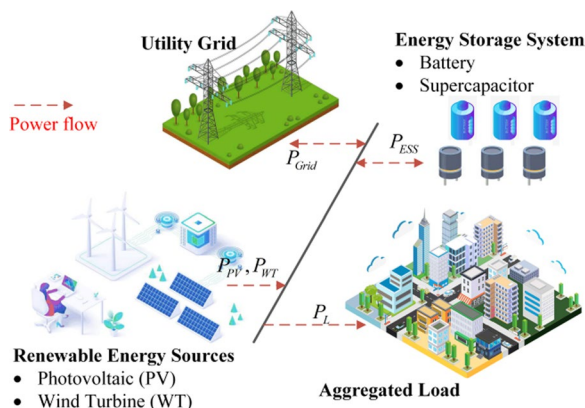


Fig. 1 Microgrid model for energy management

2.2 Renewable energy sources

PV panels and WTs are deployed as generation units. PV generation varies dramatically with weather conditions, especially for passing clouds. Similar, wind power depends on wind speed that varies from day to day, and from season to season. The power generation uncertainty makes it difficult to deal with the instantaneous supply–demand balance. The objective of the proposed multi-timescale coordinated method is to dispatch different storage devices based on their characteristics to deal with the instantaneous fluctuations of RES. The PV and WT hourly predicted power generation data are from [9], which are used in the first stage. In order to simulate the uncertain fluctuations of RES, a 10% forecast error is added on the predicted data in the second stage.

2.3 Hybrid energy storage systems (HESSs)

The microgrid has two kinds of ESS: battery and supercapacitor. The battery has high energy density to store energy, while the supercapacitor has high power density and can rapidly respond to charging/discharging events. Because of their different characteristics, the battery and supercapacitor should be scheduled for different objectives, i.e., the battery is scheduled for economic dispatch over a longer timescale while the supercapacitor maintains instantaneous power balance. Particularly, the degradation cost during operation is considered in this paper, where the models are presented in the following sections.

2.3.1 Battery degradation model

Degradation of life is important for battery operation, and is visible in two main aspects, i.e., the aging of cycle life and the reduction of capacity. The primary determinant of life-cycle and capacity is the depth of discharge (DOD), which is defined as the energy that is discharged from a fully charged battery, divided by the battery capacity. The curve of battery lifecycles under different DODs, L_B , can be fitted by the formula in [9]:

$$L_B(DOD) = a \times DOD^{-b} \times e^{-c \cdot DOD} \quad (2)$$

where a , b , c are curve-fitting coefficients.

Then the degradation cost $C_{BDC}(t)$ related to the discharging/charging events can be expressed as:

$$C_{BDC}(t) = \frac{C_B^R \cdot |P_B(t)| \Delta t}{2L_B \cdot E_{cap}^B(t) \cdot DOD \cdot \eta_c^B \cdot \eta_d^B} \quad (3)$$

where C_B^R is the battery replacement cost, and η_c^B and η_d^B are the charging and discharging efficiencies, respectively. It is worth mentioning that the degradation cost of charging events equals that of discharging events. The battery capacity will depreciate after this cycling event as:

$$E_{cap}^B(t + \Delta t) = E_{cap}^B(t) - \frac{E_{rated}^B}{L_B} \Delta t \quad (4)$$

where E_{rated}^B is the rated capacity of the battery.

2.3.2 Supercapacitor degradation model

The predominant advantage of a supercapacitor over a battery is that the supercapacitor can weather several tens of thousands of discharge/charge cycles under very high current. Degradation of its lifetime centres on the increase of equivalent series resistance and the reduction of capacity. Reference [24] indicates that the aging of a supercapacitor is closely related to the evaporation rate of electrolytes, while the temperature and voltage are two principal factors influencing the aging rate.

A supercapacitor can work for its estimated lifetime under normal operating conditions (i.e., within the proper temperature and voltage ranges). Therefore, the supercapacitor degradation cost can be treated as a linear function of DOD per charging/discharging event [10]. Considering a time interval $[t, t + \Delta t]$ and given the estimated supercapacitor lifetime L_{SC} and the replacement cost C_{SC}^R , the degradation cost can be expressed as:

$$C_{SCDC}(t) = \frac{C_{SC}^R \cdot \Delta t}{L_{SC}} \quad (5)$$

As (5) shows, the cost is linear over time if used properly in a microgrid. Therefore, a supercapacitor is suitable for dealing with instantaneous power imbalance.

3 Multi-timescale coordinated two-stage operation of microgrid with HESS

In this section, a two-stage multi-timescale coordinated method considering the complementary characteristics of battery and supercapacitor is developed. The first stage addresses a deterministic optimization problem using the hourly predicted data, and then corrective action is taken in the second stage to compensate for the first-stage decision after the uncertainty is realized.

Figure 2 illustrates the framework of the proposed DRL-based multi-timescale coordinated microgrid operation, where T_f and T_s are the lengths of the prediction horizons for the first stage and second stage, respectively. The first stage consists of a nonlinear model predictive control framework with a time horizon $t_f \in \{1, 2, \dots, T_f\}$, while the second stage is a Markov decision process solved by a DRL algorithm (Soft Actor-Critic) with a time horizon $t_s \in \{1, 2, \dots, T_s\}$, using the reference values provided by the first stage. For both stages, the power balance constraint shown in (1) needs to be met at all times. In addition, the following constraints also need to be satisfied:

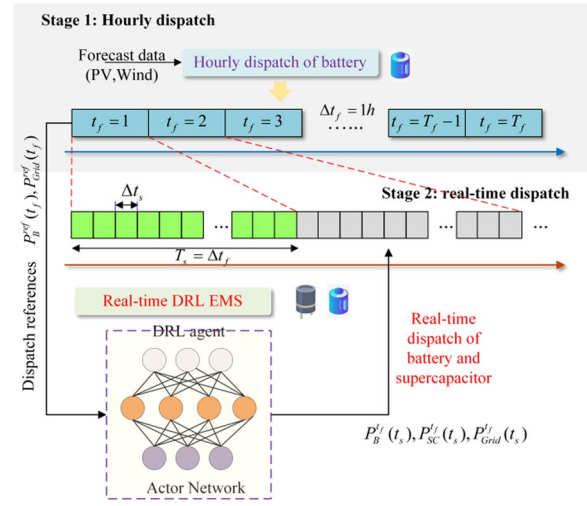


Fig. 2 Framework of the proposed DRL based multi-timescale coordinated microgrid operation

$$P_{Grid}^{\min} \leq P_{Grid}(t) \leq P_{Grid}^{\max}, t \in \{t_f, t_s\} \quad (6)$$

$$P_B^{\min} \leq P_B(t) \leq P_B^{\max}, t \in \{t_f, t_s\} \quad (7)$$

$$P_{SC}^{\min} \leq P_{SC}(t_s) \leq P_{SC}^{\max} \quad (8)$$

$$SOC_B^{\min} \leq SOC_B(t) \leq SOC_B^{\max}, t \in \{t_f, t_s\} \quad (9)$$

$$SOC_{SC}^{\min} \leq SOC_{SC}(t) \leq SOC_{SC}^{\max}, t \in \{t_f, t_s\} \quad (10)$$

Equation (6) is the power limit constraint of buying/selling electricity from/to utility grid, while (7), (8) are the charging or discharging power limits of the battery and supercapacitor, respectively. Equations (9), (10) are the state of charge (SOC) limits required to avoid the battery and supercapacitor over-charging or over-discharging, respectively.

State dynamics of the battery and supercapacitor are depicted as (11) in terms of charging/discharging power [9]:

$$E^B(t + \Delta t) = \begin{cases} E^B(t) - P_B(t) \Delta t \cdot \eta_c^B, & P_B(t) \leq 0 \\ E^B(t) - P_B(t) \Delta t / \eta_d^B, & P_B(t) > 0 \end{cases} \quad (11)$$

$$E^{SC}(t_s + \Delta t_s) = \begin{cases} E^{SC}(t_s) - P_{SC}(t_s) \Delta t_s \cdot \eta_c^{SC}, & P_{SC}(t_s) \leq 0 \\ E^{SC}(t_s) - P_{SC}(t_s) \Delta t_s / \eta_d^{SC}, & P_{SC}(t_s) > 0 \end{cases} \quad (12)$$

where $t \in \{t_f, t_s\}$ and $\Delta t \in \{\Delta t_f, \Delta t_s\}$, $E^B(t)$ and $E^{SC}(t)$ are the remaining energy in the battery and supercapacitor at time t respectively, and $SOC_{B/SC} = E^{B/SC}/E_{cap}^{B/SC}$. Since the supercapacitor is scheduled only at the second stage, the state dynamics in (12) and constraint in (8) are only considered in the second stage with a short timescale.

3.1 Mathematical model for first stage

The objective of the first stage is to optimize the decision variables $\{P_{Grid}^f(t_f), P_B^f(t_f)\}$ with hourly predicted data, including load, PV, and WT power generation, to minimize the operational cost. The objective function is expressed as:

$$\sum_{t_f=1}^{T_f} C_{Grid}^f(t_f) + C_B(t_f) \quad (13)$$

where C_{Grid}^f is the cost of buying electricity from the utility grid, which can be expressed as:

$$C_{Grid}^f(t_f) = \begin{cases} c_g^{buy}(t_f) \cdot P_{Grid}^f(t_f) \Delta t_f, & P_{Grid}^f(t_f) > 0 \\ c_g^{sell}(t_f) \cdot P_{Grid}^f(t_f) \Delta t_f, & P_{Grid}^f(t_f) \leq 0 \end{cases} \quad (14)$$

where $c_g^{buy}(t_f)$ is the buying price in the market at time t_f and $c_g^{sell}(t_f)$ is the selling price. When $P_{Grid}^f > 0$, it means the microgrid buys electricity from the utility grid, and when $P_{Grid}^f < 0$ the microgrid sells electricity to the grid. t_f is the battery operation cost, given as:

$$C_B(t_f) = C_{BDC}(t_f, DOD(t_f)) - [1 - \delta(t_f)] \cdot C_{BDC}(t_f, DOD(t_f - 1)) \quad (15)$$

where $\delta(t)$ is an auxiliary function to indicate the state transition on the charging and discharging events in two consecutive time intervals, as:

$$\delta(t_f) = \begin{cases} 0, & P_B^f(t_f) \cdot P_B^f(t_f - 1) > 0 \\ 1, & P_B^f(t_f) \cdot P_B^f(t_f - 1) \leq 0 \end{cases} \quad (16)$$

The depth of discharge, $DOD(t_f)$, is calculated based on the accumulated energy before the cycling event changes as:

$$DOD(t_f) = \frac{E_a^B(t_f)}{E_{cap}^B(t_f)} \quad (17)$$

where $E_a^B(t_f)$ is the accumulated energy before the cycling event changes, formulated as:

$$E_a^B(t_f) = [1 - \delta(t_f)] \cdot E_a^B(t_f - 1) + P_B^f(t_f) \Delta t_f \quad (18)$$

Based on the above equations, the optimization problem in the first stage is summarized as:

$$\begin{aligned} \min \sum_{t_f=1}^{T_f} C_{Grid}^f(t_f) + C_B(t_f) \\ s.t. (1) - (4), (6), (7), (9), (11), (14) - (18) \end{aligned} \quad (19)$$

3.2 Mathematical model for second stage

The second stage aims to modify the decisions (i.e., battery output power and utility regulation) from the first stage and make a schedule for the supercapacitor so that variations caused by RES uncertainties can be minimized. The decision variables are $\{P_{Grid}^{s,t_f}(t_s), P_B^{s,t_f}(t_s), P_{SC}^{s,t_f}(t_s)\}$ for each first-stage time slot $[t_f, t_f + \Delta t_f]$, and they should satisfy (6)-(8). This stage is modeled as a Markov decision process. To apply the DRL algorithm and achieve the real-time decision-making, several necessary components (i.e., observations, actions, reward function) should first be defined.

3.2.1 Observations

For the second stage, the environment is the microgrid, and therefore the state of the environment (usually termed as observations) at t_s in the time slot $[t_f, t_f + \Delta t_f]$ can be expressed as:

$$s(t_s) = \{P_L^{net}(t_s), SOC_B(t_s), SOC_{SC}(t_s)\} \quad (20)$$

where $SOC_B(t_s), SOC_{SC}(t_s)$ are the current states of charge of the battery and supercapacitor, respectively, and satisfy (9) and (10). $P_L^{net}(t_s)$ is the net load at current time step, given by:

$$P_L^{net}(t_s) = P_L(t_s) - P_{PV}(t_s) - P_{WT}(t_s) \quad (21)$$

3.2.2 Actions

The actions performed by the DRL agent are corrective items for the first stage decisions and the output power of the supercapacitor, shown as:

$$a(t_s) = \{\Delta P_{Grid}(t_s), \Delta P_B(t_s), P_{SC}(t_s)\} \quad (22)$$

The relations between the decision variables $\{P_{Grid}^{s,t_f}(t_s), P_B^{s,t_f}(t_s), P_{SC}^{s,t_f}(t_s)\}$ and actions are summarized as:

$$P_{Grid}^{s,t_f}(t_s) = P_{Grid}^f(t_f) + \Delta P_{Grid}(t_s) \quad (23)$$

$$P_B^{s,t_f}(t_s) = P_B^f(t_f) + \Delta P_B(t_s) \quad (24)$$

$$P_{SC}^{s,t_f}(t_s) = P_{SC}(t_s) \quad (25)$$

From (23)–(25), it can be seen that the actions generated by the second stage are corrective terms for the decisions made in the first stage. The first stage is a deterministic optimization process without considering the uncertainties of RES, and therefore an additional second stage is introduced so that the trained DRL agent can make real-time decisions after the uncertainties are realized.

3.2.3 Reward function

The state transition follows the rule stated in (11) and (12). As mentioned, the first stage provides reference values for the second stage and the realized uncertainties are dealt with by the supercapacitor. Therefore, penalties for deviations from the reference values can be defined as:

$$c_g^p(t_s) = [\Delta P_{Grid}(t_s)]^2 \quad (26)$$

$$c_B^p(t_s) = [\Delta P_B(t_s)]^2 \quad (27)$$

where $c_g^p(t_s)$ is the penalty cost of deviation of the power from/to utility grid, and $c_B^p(t_s)$ is the penalty cost of deviation of the battery power. Apart from these two penalties, the SOC of the supercapacitor at the end of each time interval $[t_s - \Delta t_s, t_s]$ should be maintained at a nominal value so that it has enough energy for the next time interval. Therefore, the penalty term that accounts for the SOC of the supercapacitor can be expressed as:

$$c_{sc}^p(t_s) = (SOC_{SC}(t_s)/100 - 1)^2 \quad (28)$$

Based on (27)–(29), the reward function is defined as:

$$r(t_s) = -c_g^p(t_s) - c_B^p(t_s) - c_{sc}^p(t_s) - C_{SCDC} \quad (29)$$

where the first three items are penalty costs, and the last is the supercapacitor degradation cost.

With the above definitions, the optimization problem in the second stage when considering the future rewards is formulated as:

$$\max \sum_{t_s=1}^{T_s} \gamma^{t_s} \cdot r(t_s)$$

$$s.t. (1), (5) - (12), (20) - (29) \quad (30)$$

where γ is the discount factor. This problem will be solved by the DRL algorithm, SAC. Since the policy is composed of deep neural networks, a completely trained agent can compute the dispatch actions within a few seconds for real-time decision-making.

3.3 SAC training process

The SAC is one of the actor-critic approaches similar to DDPG and Twin delayed DDPG (TD3) but has better performance for continuous control problems. In the actor-critic architecture, the actor uses the policy gradient to find the optimal policy while the critic evaluates the policy produced by the actor. Instead of using only the Q function, the critic in the SAC takes both the Q function and the value function to evaluate the policy. This can stabilize the training and provide convenience to train simultaneously with the other networks. There are three networks in the SAC, one actor network (policy network) to find the optimal policy, and two critic networks (a value network and a Q network) to evaluate the policy. They will be explained in detail in the following sections.

3.3.1 Critic network

The critic in the SAC uses both the value function and the Q function to evaluate the learned policy, and the two functions are approximated by two separate deep neural networks. The value network is denoted by V , the parameter of the network is ψ and the parameter of the target value network is ψ' . Therefore, $V_\psi(s)$ is the value function (state value) approximated by a neural network parameterized by ψ . In the entropy-based reinforcement learning framework, the theoretical state value V is calculated as:

$$V(s) = E_\pi[Q(s, a) - \alpha \log \pi(a|s)] \quad (31)$$

where $Q(s, a)$ is the state-action value function (Q-value), π is the stochastic policy that maps the state to a probability distribution over the action space, $\alpha \log \pi(a|s)$ is the entropy term, and a_i is the entropy weight, which is adjusted dynamically during the training process. Two neural networks $Q_{\theta 1}$ and $Q_{\theta 2}$ are used to compute the Q value, and this can avoid overestimation during the training process [25]. The policy is expressed by the actor network (policy network) π_ϕ parameterized by ϕ . Therefore, the loss function of the value network can be defined as:

$$L_V(\psi) = \frac{1}{N} \sum_i [\min_{j=1,2} Q_{\theta_j}(s_i, a_i) - \alpha \log \pi_\phi(a_i|s_i) - V_\psi(s_i)]^2 \quad (32)$$

where N is the number of state transitions sampled from replay buffer D , i is the index of sample, s_i and a_i are state observation and action defined in (20) and (22), respectively. The value network parameter ψ is updated by gradient descent, as:

$$\psi = \psi - \lambda \nabla_{\psi} L_V(\psi) \quad (33)$$

where λ is the learning rate. Then the target value network is updated by soft replacement, as:

$$\psi' = \tau\psi + (1 - \tau)\psi' \quad (34)$$

where τ is the soft replacement coefficient that is set to 0.001.

For the two Q networks, Q_{θ_1} and Q_{θ_2} , the loss functions are defined as:

$$L_Q(\theta_1) = \frac{1}{N} \sum_i [r_i + \gamma V_\psi(s'_i) - Q_{\theta_1}(s_i, a_i)]^2 \quad (35)$$

$$L_Q(\theta_2) = \frac{1}{N} \sum_i [r_i + \gamma V_\psi(s'_i) - Q_{\theta_2}(s_i, a_i)]^2 \quad (36)$$

where r_i is the immediate reward defined in (29), γ is the discount factor that gives importance to the immediate reward or future rewards. s'_i is the new state after the action a_i is performed. Then θ_1 and θ_2 are updated by the following rules:

$$\theta_1 = \theta_1 - \lambda \nabla_{\theta_1} L_Q(\theta_1) \quad (37)$$

$$\theta_2 = \theta_2 - \lambda \nabla_{\theta_2} L_Q(\theta_2) \quad (38)$$

3.3.2 Critic network

The actor network, also named policy network, is parameterized by ϕ . The loss function is defined as:

$$L_{\pi}(\phi) = \frac{1}{N} \sum_i [\min_{j=1,2} Q_{\theta_j}(s_i, a_i) - \alpha \log \pi_{\phi}(a_i | s_i)]^2 \quad (39)$$

Then the actor network parameter can be updated by:

$$\phi = \phi + \lambda \nabla_{\phi} L_{\pi}(\phi) \quad (40)$$

3.4 Algorithm

The proposed DRL based method for multi-timescale coordinated operation of microgrid is summarized in Algorithm 1.

Algorithm 1: DRL based multi-timescale coordinated microgrid operation

Input Data: real-time electricity price, PV and WT power generation

Output: Optimal scheduling policy

1. Initialize the value network parameter ψ , the Q network parameters θ_1 and θ_2 , and the actor network parameter ϕ ;
2. Initialize the target value network parameter $\psi' = \psi$;
3. Initialize the replay buffer D ;
4. **for** episode = 1 to M
5. **for** $t_f = 1$ to T_f
6. solve the problem (20) and get $P_{Grid}^f(t_f), P_B^f(t_f)$;
7. **for** $t_s = 1$ to T_s
8. get the observation and select an action $a = \pi_\phi(s)$;
9. perform the action a , move to the next state s' , get
nnnnnnnnnnnnthe reward r ;
10. store the transition $\langle s, a, s', r \rangle$ into the replay buffer;
11. randomly sample a minibatch with N transitions
nnnnnnnnnnnnfrom D ;
12. compute the loss function of value network $L_v(\psi)$;
13. update the value network according to (34);
14. compute the loss function of Q network
nnnnnnnnnnnn $L_Q(\theta_1), L_Q(\theta_2)$ n;
15. update the Q network according to (38) and (39);
16. compute the loss function of actor network $L_\pi(\phi)$;
17. update the actor network according to (41);
18. update the target value network according to (35);
19. **end**
20. **end**
21. **end**

4 Case studies

4.1 Simulation settings

In this section, the DRL based method for multi-time-scale coordinated operation of microgrids is demonstrated. The model shown in Sect. 3 and the SAC algorithm are implemented in MATLAB. The optimization problem in the first stage is solved by Gurobi [26],

Table 1 Energy storage device parameters

Device	Capital Cost (\$/kWh)	Capacity (kWh)	Power Limit (kW)	SOC ^{min}	SOC ^{max}
Battery	600	12	4	10%	90%
Supercapacitor	3600	1	10	0%	100%

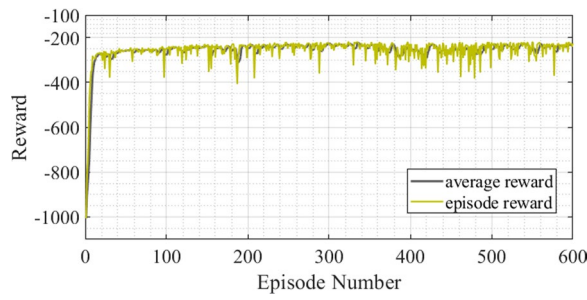


Fig. 3 Episode and average rewards during the training process

and the DRL training process is conducted by the MATLAB Reinforcement Learning Toolbox [27].

The simulation parameters are listed in Table 1. The scheduling horizon is set as 48 h, the length of the time interval in the first stage is 1 h and the time interval in the second stage is set to 5 min. It is worth mentioning that although in this paper the time interval is set as 5 min for simplicity, the proposed method can adapt to a shorter one. The data about the power of load, PV and WT can be found in [9], while the uncertainties of the PV and WT power generation in the second stage are expressed by a 10% forecast error. The real-time electricity price data are obtained from [23], and the charging/discharging efficiencies of the battery and supercapacitor are set as 0.95 and 0.92, respectively. For the DRL training process, the DRL parameters, including the iteration number, learning rate λ , discount factor γ , size of replay buffer and soft replacement coefficient τ , are set as 600, 3×10^{-4} , 0.99, 1000, and 0.001, respectively. Figure 3 shows the reward value of each training episode of the SAC agent. As shown, episode reward is the total reward in the scheduling horizon 48 h, and average reward is the average of rewards of all finished episodes. After several hundreds of training episodes, episode and average rewards converge to a maximum value, and it can be seen that the SAC algorithm converges very quickly. After the training process, entropy weight α is set to 0.2296 and the DRL agent with the optimal policy is implemented for real-time decision-making.

4.2 Numerical results

The results of optimal dispatch based on the real-time pricing scheme computed by the proposed two-stage framework are shown in Figs. 4 and 5. Figure 4 is the results of the first stage without considering the uncertainties of PV and WT power generation, whereas Fig. 5 is the results of the second stage considering the uncertainties. As shown, the battery is rapidly discharged at hours 16, 20 and 37 when the electricity price is high,

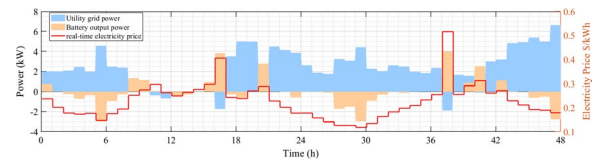


Fig. 4 First-stage economic dispatch without considering the uncertainty

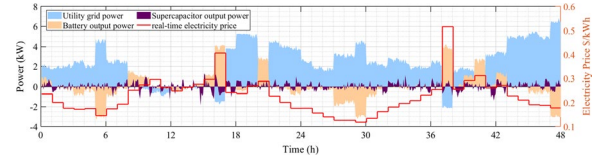


Fig. 5 Second-stage economic dispatch considering the PV and WT uncertainties (10% forecast error)

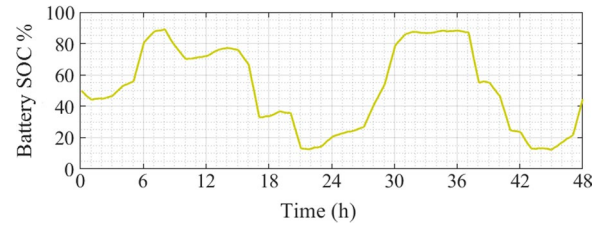


Fig. 6 SOC of battery during dispatch horizon

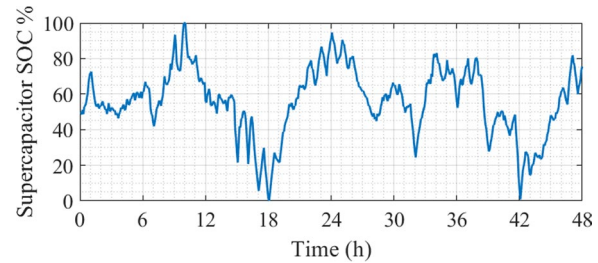


Fig. 7 SOC of supercapacitor during dispatch horizon

while when the electricity price is lower from hours 0–8 and 21–35 the microgrid buys a large amount of electricity from the main grid to satisfy demand and charge the battery. The average computation time of the first stage is 18.85 s. From Fig. 5, it can be seen that the supercapacitor is scheduled to rapidly respond to the instantaneous fluctuations of PV and WT power generation. The computation time of a step in the second stage is 0.86 s. Figures 6 and 7 show the SOC of the battery and supercapacitor during the scheduling horizon respectively. These present the periodic features as expected. As seen, the SOC curve of the battery is much smoother than

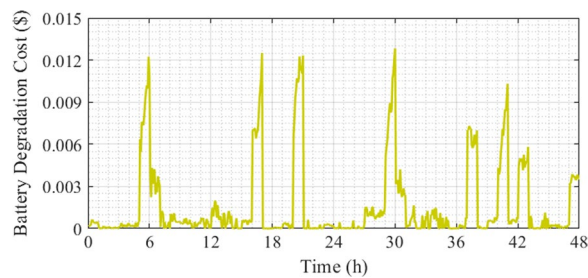


Fig. 8 Battery degradation cost during dispatch horizon

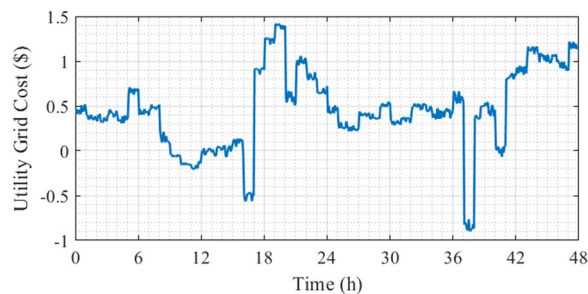


Fig. 9 Utility grid regulation cost during dispatch horizon

that of the supercapacitor since the output power of the supercapacitor is closely related to the variations of RES.

Figures 8 and 9 provide the overview of battery degradation cost and utility grid regulation cost over the scheduling horizon, respectively. Comparing Figs. 8 and 5, it can be seen that the degradation cost is high when the battery is quickly charged or discharged. It can also be seen from Figs. 9 and 5 that the profile of the utility grid regulation cost is closely related to the electricity price. Table 2 gives the detailed dispatch results. As seen, the total operation cost is \$264.2974 and the battery degradation cost is \$0.9222. After the DRL agent is completely trained, it takes 0.1563 s to make the decisions for each step. This is sufficiently quick for real-time operation.

5 Conclusions

In this study, a DRL based method for multi-timescale coordinated microgrid operation is developed. The first stage makes an initial decision irrespective of the uncertainties to minimize the operation and degradation costs, while the second stage aims to generate corrective actions

to compensate for the first-stage decision after the uncertainty is realized. The first stage is a non-linear model predictive control problem while the second stage is modeled as a Markov decision process solved by an SAC. Since the battery has high energy intensity while the supercapacitor has high power intensity, the battery is scheduled to store a large amount of energy over a long timescale while the supercapacitor works for instantaneous power balancing. The training results of the SAC agent show that this method can stabilize the performance well and converges to the optimal policy quickly. For online application, the agent can adjust the power bought from the main grid, battery power, and supercapacitor power in real-time to maintain the power balance under the fluctuations of RES.

In summary, this paper establishes a feasible data-driven multi-timescale coordinated microgrid operation method. In future work, the problem in the first stage will be further analyzed and solved by the DRL algorithm to form a multi-agent DRL-based coordinated framework. In addition, more types of energy storage device will be considered.

Author information

Chunchao Hu: male, PHD Candidate and Engineer. Major in power system protective, control and automation, and computational intelligence in operation optimization of power systems.

Zexiang Cai: male, PHD and Professor. Major in power system stability and control.

Yanxu Zhang: male, PHD and Engineer. Major in power system protective, control and automation, and computational intelligence in operation optimization of power systems.

Rudai Yan: male, PHD Candidate. Major in distributed control and optimization in microgrids, and data analytics in the control and operation of power systems.

Yu Cai: male, PHD. Major in power system optimal operation and control and smart grid.

Bowei Cen: male, Master Degree Candidate. Major in power system optimal operation and control and smart grid.

Acknowledgements

First of all, I would like to extend my sincere gratitude to my tutor, Professor Zexiang Cai, for his instructive advice on my thesis. I am also deeply indebted to the assistance from all the lab-mates of South China University of Technology and the cooperation with China Southern Power Grid Electrical Technology Co., LTD and Nanyang Technological University. Finally, I would like to express my heartfelt thankfulness for Guangdong Provincial Key Laboratory of New Technology for Smart Grid Funded Projects which provides the experimental support to my work.

Author contributions

CH: Conceptualization, Methodology, Data Curation, Writing-Original Draft. ZC: Funding acquisition, Methodology. YZ: Writing-Review & Editing, Formal analysis, Software. RY: Resources, Investigation, Software, Visualization. YC: Supervision, Formal analysis, Software. BC: Supervision, Writing-Review. All authors read and approved the final manuscript.

Funding

The work was supported by Guangdong Provincial Key Laboratory of New Technology for Smart Grid Funded Project under Grant No. 2020b1212070025.

Availability of data and materials

All the data and materials that are required can be shared by contacting the corresponding author.

Table 2 Numerical results

Terms	Cost(\$)
Total operation cost	264.2974
Average operation cost (5 min)	0.4588
Battery degradation cost	0.9222

Declarations

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author details

¹School of Electric Power, South China University of Technology, Guangzhou, China. ²China Southern Power Grid Electrical Technology Co., LTD, Guangzhou, China. ³Guangdong Provincial Key Laboratory of New Technology for Smart Grid, Guangzhou, China. ⁴School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore. ⁵School of Electrical and Electronic Engineering, South China University of Technology, Guangzhou 510640, China.

Received: 19 February 2022 Accepted: 22 July 2022

Published online: 05 August 2022

References

- Olivares, D. E., Mehrizi-Sani, A., Etemadi, A. H., Cañizares, C. A., Iravani, R., Kazerani, M., & Hatziargyriou, N. D. (2014). Trends in microgrid control. *IEEE Transactions on Smart Grid*, 5(4), 1905–1919. <https://doi.org/10.1109/TSG.2013.2295514>
- Zhang, C., Xu, Y., Dong, Z. Y., & Ma, J. (2017). Robust operation of microgrids via two-stage coordinated energy storage and direct load control. *IEEE Transactions on Power Systems*, 32(4), 2858–2868. <https://doi.org/10.1109/TPWRS.2016.2627583>
- Guerrero, J. M., Vasquez, J. C., Matas, J., De Vicuña, L. G., & Castilla, M. (2010). Hierarchical control of droop-controlled AC and DC microgrids—a general approach toward standardization. *IEEE Transactions on Industrial Electronics*, 58(1), 158–172.
- Jing, W., Lai, C. H., Wong, S. H. W., & Wong, M. L. D. (2017). Battery-supercapacitor hybrid energy storage system in standalone DC microgrids: A review. *IET Renewable Power Generation*, 11(4), 461–469.
- Shen, J., & Khaligh, A. (2015). A supervisory energy management control strategy in a battery/ultracapacitor hybrid energy storage system. *IEEE Transactions on Transportation Electrification*, 1(3), 223–231.
- Hredzak, B., Agelidis, V. G., & Demetriades, G. D. (2013). A low complexity control system for a hybrid dc power source based on ultracapacitor–lead–acid battery configuration. *IEEE Transactions on Power Electronics*, 29(6), 2882–2891.
- Suganthi, L., Iniyar, S., & Samuel, A. A. (2015). Applications of fuzzy logic in renewable energy systems—a review. *Renewable and Sustainable Energy Reviews*, 48, 585–607.
- Hredzak, B., Agelidis, V. G., & Jang, M. (2013). A model predictive control system for a hybrid battery-ultracapacitor power source. *IEEE Transactions on Power Electronics*, 29(3), 1469–1479.
- Ju, C., Wang, P., Goel, L., & Xu, Y. (2018). A two-layer energy management system for microgrids with hybrid energy storage considering degradation costs. *IEEE Transactions on Smart Grid*, 9(6), 6047–6057. <https://doi.org/10.1109/TSG.2017.2703126>
- Arabali, A., Ghofrani, M., Etezadi-Amoli, M., Fadali, M. S., & Baghzouz, Y. (2012). Genetic-algorithm-based optimization approach for energy management. *IEEE Transactions on Power Delivery*, 28(1), 162–170.
- Nafisi, H., Agah, S. M. M., Abyaneh, H. A., & Abedi, M. (2015). Two-stage optimization method for energy loss minimization in microgrid based on smart power management scheme of PHEVs. *IEEE Transactions on Smart Grid*, 7(3), 1268–1276.
- Hu, W., Wang, P., & Gooi, H. B. (2016). Toward optimal energy management of microgrids via robust two-stage optimization. *IEEE Transactions on Smart Grid*, 9(2), 1161–1174.
- Xiao, J., Wang, P., & Setyawati, L. (2015). Hierarchical control of hybrid energy storage system in DC microgrids. *IEEE Transactions on Industrial Electronics*, 62(8), 4915–4924.
- Mendis, N., Muttaqi, K. M., & Perera, S. (2014). Management of battery-supercapacitor hybrid energy storage and synchronous condenser for isolated operation of PMSG based variable-speed wind turbine generating systems. *IEEE Transactions on Smart Grid*, 5(2), 944–953.
- Ghiassi-Farokhfal, Y., Rosenberg, C., Keshav, S., & Adjaho, M.-B. (2016). Joint optimal design and operation of hybrid energy storage systems. *IEEE Journal on Selected Areas in Communications*, 34(3), 639–650.
- Ji, Y., Wang, J., Xu, J., Fang, X., & Zhang, H. (2019). Real-time energy management of a microgrid using deep reinforcement learning. *Energies*, 12(12), 2291.
- Hua, H., Qin, Y., Hao, C., & Cao, J. (2019). Optimal energy management strategies for energy Internet via deep reinforcement learning approach. *Applied Energy*, 239, 598–609. <https://doi.org/10.1016/j.apenergy.2019.01.145>
- Du, Y., & Li, F. (2019). Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning. *IEEE Transactions on Smart Grid*, 11(2), 1066–1076.
- Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. 35th International Conference on Machine Learning, ICML 2018, 5, 2976–2989.
- Yan, Z., & Xu, Y. (2020). A multi-agent deep reinforcement learning method for cooperative load frequency control of a multi-area power system. *IEEE Transactions on Power Systems*, 35(6), 4599–4608.
- Wang, W., Yu, N., Gao, Y., & Shi, J. (2019). Safe off-policy deep reinforcement learning algorithm for Volt-VAR control in power distribution systems. *IEEE Transactions on Smart Grid*, 11(4), 3008–3018.
- Fakhrazari, A., Vakiliadian, H., & Choobineh, F. F. (2014). Optimal energy scheduling for a smart entity. *IEEE Transactions on Smart Grid*, 5(6), 2919–2928.
- Energy Market Company (EMCSG) Available online: <https://www.emcsg.com/marketdata>.
- Alcicek, G., Gualous, H., Venet, P., Gallay, R., & Miraoui, A. (2007). Experimental study of temperature effect on ultracapacitor ageing. In 2007 European conference on power electronics and applications (pp. 1–7). IEEE.
- Ravichandiran, S. (2018). Hands-on reinforcement learning with Python: master reinforcement and deep reinforcement learning using OpenAI gym and tensorflow. Packt Publishing Ltd.
- Gurobi Optimization. Available online: <http://www.gurobi.com>.
- Matlab Reinforcement Learning Toolbox. Available online: <https://in.mathworks.com/products/reinforcement-learning.html>.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)