A real-time energy management strategy of flexible smart traction power supply system based on deep Q-learning
Ying, Yichen; Tian, Zhongbei; Wu, Mingli; Liu, Qiujiang; Tricoli, Pietro

License:
Other (please specify with Rights Statement)

Document Version
Peer reviewed version

Citation for published version (Harvard):
Ying, Y, Tian, Z, Wu, M, Liu, Q & Tricoli, P 2024, 'A real-time energy management strategy of flexible smart traction power supply system based on deep Q-learning', IEEE Transactions on Intelligent Transportation Systems.

Link to publication on Research at Birmingham portal

Publisher Rights Statement:
© 2024 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

General rights
Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

• Users may freely distribute the URL that is used to identify this publication.
• Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
• Users may use extracts from the document in line with the concept of ‘fair dealing’ under the Copyright, Designs and Patents Act 1988 (?
• Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.
A real-time energy management strategy of flexible smart traction power supply system based on deep Q-learning

Yichen Ying, Member, IEEE, Zhongbei Tian, Member, IEEE, Mingli Wu, Member, IEEE, Qiujiang Liu, Member, IEEE, and Pietro Tricoli, Senior Member, IEEE

Abstract—Due to the high degree of controllability of the flexible smart traction power supply system (FSTPSS), day-ahead energy management strategy (DAEMS) was developed to optimize the power flow of the FSTPSS. However, the use of DAEMS is not based on real-time information. For FSTPSS, without real-time information, it cannot solve the problem of planning deviation caused by the real-time fluctuation of uncertain loads or sources. Therefore, this paper proposes a real-time energy management strategy (REMS) which is based on the real-time information to address the problem of planning deviation. REMS is implemented by LSTM and deep Q learning algorithm, where LSTM predicts uncertain loads or sources, and the deep Q-learning controls the operation of FSTPSS based on real-time predicted state. The proposed strategy is validated with the power flow simulation model of TPSS and the real measured data. The simulation results verify the necessity and superiority of the proposed method.

Index Terms—real-time information, energy management, planning deviation, deep Q-learning, traction power supply system

I. INTRODUCTION

The electric power for electrified railways comes from TPSS. The traditional TPSS has the advantages of simple structure and low cost, but it also has a lot of problems: (1) Great impact of traction load on power grid and TPSS. From the statistics of [1], the maximum active power of the traction station can reach 40MW; (2) Three-phase unbalance due to transformer construction. The magnitude of the negative sequence current can be distributed between tens of amperes to more than one hundred amperes and this will cause the negative sequence current at the common connection point to be exceeded. [2, 3]; (3) Insufficient utilization of train regenerative braking energy. In South Nanjing substation, the daily regenerative braking energy sent back to the substation can reach 60MWh [4]. (4) no interface for renewable energy access; (5) overvoltage and ferromagnetic resonance caused by electrical phase separation [5].

The new idea of TPSS topology has put forward some possibilities for solving the above problems. In [6], a novel TPSS topology based on the RPC connecting both sides of the neutral section (NS) was proposed, which enables the regenerative braking energy to flow on both sides of NS, thereby improving the utilization efficiency of the regenerative braking energy. In [7], an AC-DC-AC substation was proposed to replace the traditional traction substation, so that the three-phase unbalance problem existing in traditional TPSS can be solved. In [8], a wayside microgrid was proposed to connect the PV scattered around the railway line. This microgrid could be directly connected to TPSS. It can solve the problem of no interface for renewable energy access. In [9], a new TPSS topology was developed based on AC-DC-AC traction substation, which connected HESS and PV to the DC side of the traction substation. And a DAEMS was proposed to optimize the whole day power flow of this novel TPSS. The various new TPSS topologies mentioned above have given us some inspiration. At the same time, we also found that when controllable power electronic devices are added to TPSS, an appropriate energy management strategy is very necessary. In [10], a two-stage robust optimization model based on combined co-phase power supply system with the integration of PV and hybrid energy storage system was proposed to handle the uncertainties of PV and traction load. The three-phase voltage unbalance constraint is taken into account in particular and a column-and-constraint generation algorithm is used to solve this robust optimization model. A railway electrical smart grid and a chance-constrained two-stage programming approach was proposed in [11] to minimize the total cost of the system under uncertain traction load. The simulation results show that compared with the traditional uncertain modeling approaches, the model provides significant improvements in reliability and sharpness features. However, all of the above energy management strategy are DAEMS. If we only use DAEMS to manage the energy of system, there will be the following problems: First, the time scale planned by DAEMS is larger, and it does not take into account the load fluctuation of smaller scale [12]. Meanwhile, since the load data of DAEMS is based on the prediction of historical data rather than the data collected in real time, it cannot revise its plan in time. When the data collected in real time is different from the historical data, the plan of DAEMS cannot be accurately implemented [13]. In addition, in the situation of real-time prediction, the time we make prediction is very close to the prediction point, many
states are difficult to mutate, so we can exclude more possible situations which could make the accuracy of the prediction higher. Hence, a real-time prediction algorithm is needed which could make prediction results more accurate. Secondly, due to the consideration of the full day energy management strategy, the power transmission losses of the system are not considered. Since the process of calculating these losses is not a linear process, it will cause certain errors.

Based on the analysis above, to implement DAEMS reliably, there will be two issues need to be addressed. One problem is real-time prediction, and another is the accurate system losses estimation and error compensation.

The first problem is a prediction problem of time series data. Nowadays, with the development of artificial intelligence (AI), instead of modeling the specific problem, we can use a general recurrent neural network (RNN) to make predictions on time series data. However, traditional RNN are usually not effective for long-sequence data prediction. Therefore, long short-term memory (LSTM) [14] which is special for long sequence prediction was proposed to handle this problem. In [15], a LSTM based hybrid ensemble learning forecasting method was proposed to predict the ultra-short-term industrial power demand. In [16], four real-world load of consumer was successfully predicted by the method based on variational mode decomposition and LSTM. In [17], a novel hybrid multitask multi-information fusion deep learning framework based on LSTM and convolutional neural network (CNN) was used to predict household short-term load. In [18], a LSTM based framework was proposed to forecast the residential load and get very accurate results. However, at present, there are few LSTM specialized for uncertain loads prediction in TPSS. Therefore, this paper designs a kind of LSTM network which fully considers the characteristics of TPSS.

The second problem is about system losses estimation and error compensation. Once the losses of the system are known, relevant control decisions can be made to compensate for these errors. The method commonly used nowadays is to build a model to estimate these errors and then make control decisions based on the estimated errors. This approach is called model predictive control (MPC) which is widely used in many studies. Naseri et al. [19] used MPC in a supercapacitor-based active stabilizer in subway TPSS to mitigate the variations of the dc-link voltage. In [20], MPC was used to stabilize the multilevel output voltage of railway traction power conditioner. Novak et al. [21] used a hierarchical MPC to manage the energy consumption in traction substation and train. In [9], a MPC method was used to compensate the prediction error of DAEMS. The accuracy of MPC method is dependent on the accuracy of the model. However, in practice, the line parameters of TPSS are often difficult to obtain accurately and may sometimes change. In this case, the prediction accuracy of MPC may decrease. Hence, many studies have turned to model-free methods which have higher flexibility to control. The model-free method can automatically estimate the model situation of the system. Therefore, the system loss estimation and the error compensation can be implemented uniformly. Model-free methods are usually implemented by some machine learning algorithms and reinforcement learning (RL) is a promising one for control decision-making. Zhou et al. [22] used a deep reinforcement learning (DRL) based optimal power flow solutions to assist power grid operators to make rapid decision. The results showed significant control ability of DRL in real-time operation. This conclusion can also be verified by the results in [23], which used deep deterministic policy gradient to implement DRL. Zhao et al. [24] proposed a knowledge-assisted deep deterministic policy gradient algorithm to implement the optimal wind farm control. The results show this method could achieve optimal control of wind farm. In [25], a DRL method which called deep Q-learning (DQN) was used to optimize the control decision of battery energy storage system in microgrid. In [26], a DQN method was used in the grid-connected vehicles charging to protect its battery health and decrease the power fluctuations. These results show the strong learning ability of DRL. But compared to the power system, TPSS has some different characteristics, such as the rapid movement and intermittent impact of the traction load. There are also some studies which use DRL to implement energy management in TPSS. In [27], a DQN-based energy management strategy was used to manage the power flow in urban rail transit TPSS. Zhu et al. [28] proposed a collaborative energy management optimization algorithm based on multiple DQN agents. The model-free method has the advantages of strong learning ability and less human intervention. After foreseeing the great potential of DRL—a promising model-free method, we decided to try to apply it in real-time energy management strategy (REMS) of FSTPSS. At present, no one has used deep Q-learning, a DRL method, for REMS for real-time compensation of day-ahead energy management strategy errors in FSTPSS. This paper hopes to apply this model-free method in FSTPSS to explore some problems that may exist in the practical application of this method. Meanwhile, FSTPSS is also a novel TPSS topology representing the development direction of the future TPSS. Therefore, this paper proposes a novel TPSS topology and its REMS based on DQ and LSTM to explore some characteristics of the novel TPSS and the feasibility of applying REMS using the model-free method in this new TPSS. The main contributions of this paper are as follows:

1). A new type of TPSS topology based on AC-DC-AC traction substation, distributed generation and HESS is proposed, which could effectively solve the problems existing in traditional TPSS and the consumption problem of renewable energy.

2). Combined with DAEMS of FSTPSS, a REMS based on LSTM and DRL is proposed, which fully considers the characteristics of the mobility load of electrified railways. This is the first time that the model-free method is used in the real-time energy management strategy of FSTPSS.

3). The neural network is used to replace the original physical simulation model to realize a completely model-free optimization algorithm for energy management.

4). A series of simulations are carried out using the measured data to verify the effectiveness and superiority of the proposed method. Meanwhile, a series of simulations are also conducted to reflect the effect of the action interval selection of DQN on its performance.

The rest of this paper is organized as follows. Section II describes the topology of FSTPSS and gives the constraint equations of FSTPSS. At the same time, objective function of REMS is given. The details of REMS based on LSTM and DQN
is presented in Section III. In Section IV, a series of case study are presented and analyzed. The conclusion is presented in Section V.

II. PROBLEM DEFINITION

A. System Topology Description

To address the problems of regenerative braking energy utilization, three-phase unbalance, great load impact, electrical phase separation, and no interface to distributed generation access, this paper proposes a new TPSS, as shown in Fig. 1.

![Fig. 1. The specific structure of FSTPSS.](image)

It uses AC-DC-AC power converters to replace single-phase transformers, thereby eliminating the problems of electrical phase separation and three-phase unbalance. Additionally, a HESS is connected to the dc-link of the AC-DC-AC converter, typically operating at few kV, and a microgrid including energy storage devices and distributed generations is connected along the line to enable railway connection to the nearby renewable power sources. This microgrid is then connected to the power distribution network, operating at 10 kV in Fig. 1.

B. Constraints of FSTPSS

For the circuit system, the most basic constraint is power balance. Therefore, the inflow and outflow of power at each electrical node should be equal. In fact, FSTPSS could be divided into four equivalent nodes: the first equivalent node is the DC side of the AC-DC-AC traction substation, the second equivalent node is the traction network, the third equivalent node is the DC bus of microgrid, and the fourth equivalent node is the transmission line from microgrid to 10 kV distribution network. So, the following four power balance equations can be obtained:

\[ P_{\text{sub-grid}} + P_{\text{sub-ac}} + P_{\text{sub-bat}} = P_{\text{sub-t}} \]  \hspace{1cm} (1)

\[ P_{\text{sub-t}} + P_{\text{dc-1}} = P_{\text{t}} \]  \hspace{1cm} (2)

\[ P_{\text{pv}} + P_{\text{dc-2}} = P_{\text{dc-1}} + P_{\text{dc-load}} \]  \hspace{1cm} (3)

\[ P_{\text{10kV grid}} + P_{\text{dc-load}} = P_{\text{10kV load}} \]  \hspace{1cm} (4)

Among them, \( P_{\text{sub-grid}} \) is the power provided by 220 kV power grid. \( P_{\text{sub-ac}} \) and \( P_{\text{sub-bat}} \) are the power given by ultra-capacitor (UC) and battery on the DC side of traction substation respectively. \( P_{\text{sub-t}} \) is the power delivered by the traction substation to traction load. \( P_{\text{t}} \) is the power consumed by the traction load. \( P_{\text{dc-2}} \) is the power delivered by the DC side of microgrid to the traction load. \( P_{\text{pv}} \) is the output power of PV panel. \( P_{\text{dc-1}} \) is the power given by the UC by the DC side of microgrid. \( P_{\text{dc-load}} \) is the power provided by the DC side of microgrid to the 10 kV distribution network. \( P_{\text{10kV load}} \) is the power of the 10 kV grid. \( P_{\text{10kV load}} \) is the load power of 10 kV distribution network.

In addition to the limitation of power balance, since FSTPSS contains many energy storage devices, its power output is also limited by the parameters of the energy storage device itself. The specific constraints are as follows:

\[ E_{\text{uc}}(t) = E_{\text{uc}}(t-1) + \eta_{\text{uc}}^\text{dis}(t) \Delta t - P_{\text{uc}}^\text{dis}(t) \Delta t / \eta_{\text{uc}}^\text{ch} \]  \hspace{1cm} (5)

\[ E_{\text{bat}}(t) = E_{\text{bat}}(t-1) + \eta_{\text{bat}}^\text{dis}(t) \Delta t - P_{\text{bat}}^\text{dis}(t) \Delta t / \eta_{\text{bat}}^\text{ch} \]  \hspace{1cm} (6)

\[ 0 \leq P_{\text{uc}}^\text{dis}(t) \leq \min((E_{\text{uc}}(t-1) - E_{\text{uc}}^\text{min}) / (\eta_{\text{uc}}^\text{dis} \cdot \Delta t), P_{\text{uc}}^\text{rated} \cdot \Delta t) \]  \hspace{1cm} (7)

\[ 0 \leq P_{\text{bat}}^\text{dis}(t) \leq \min((E_{\text{bat}}(t-1) - E_{\text{bat}}^\text{min}) / (\eta_{\text{bat}}^\text{dis} \cdot \Delta t), P_{\text{bat}}^\text{rated} \cdot \Delta t) \]  \hspace{1cm} (8)

\[ 0 \leq P_{\text{uc}}^\text{ch}(t) \leq \min((E_{\text{uc}}(t-1) - E_{\text{uc}}^\text{max}) / (\eta_{\text{uc}}^\text{ch} \cdot \Delta t), P_{\text{uc}}^\text{rated} \cdot \Delta t) \]  \hspace{1cm} (9)

\[ 0 \leq P_{\text{bat}}^\text{ch}(t) \leq \min((E_{\text{bat}}(t-1) - E_{\text{bat}}^\text{max}) / (\eta_{\text{bat}}^\text{ch} \cdot \Delta t), P_{\text{bat}}^\text{rated} \cdot \Delta t) \]  \hspace{1cm} (10)

where \( E_{\text{uc}} \) and \( E_{\text{bat}} \) are the remaining energy of UC and battery. \( P_{\text{uc}}^\text{dis} \) and \( P_{\text{bat}}^\text{dis} \) are the charge and discharge power of UC. \( P_{\text{uc}}^\text{ch} \) and \( P_{\text{bat}}^\text{ch} \) are the charge and discharge power of the battery. \( \eta_{\text{uc}}^\text{dis} \) and \( \eta_{\text{bat}}^\text{dis} \) indicate the charge and discharge efficiency of UC. \( \eta_{\text{uc}}^\text{ch} \) and \( \eta_{\text{bat}}^\text{ch} \) present the charge and discharge efficiency of the battery. \( \Delta t \) is the interval of calculation time. \( E_{\text{uc}}^\text{max} \) and \( E_{\text{uc}}^\text{min} \) are the rated capacity of battery and UC respectively. \( SOC_{\text{uc}}^\text{max} \) and \( SOC_{\text{uc}}^\text{min} \) are the minimum and maximum values for the state of charge of UC respectively. \( SOC_{\text{bat}}^\text{max} \) and \( SOC_{\text{bat}}^\text{min} \) are the minimum and maximum values for the state of charge of the battery respectively.

Moreover, the converters between each connection node have an upper capacity limit, so it is necessary to add the constraints of each converter itself, as shown below:

\[ S_{\text{converter}}^\text{rated} < S_{\text{converter}} < S_{\text{converter}}^\text{rated} \]  \hspace{1cm} (11)

where \( S_{\text{converter}} \) is the power passing through converter \( i \). \( S_{\text{converter}}^\text{rated} \) is the rated power of converter \( i \).

C. Day-ahead Energy Management Strategy

Based on the above constraints and equations, DAEMS focuses on minimizing the total cost of daily operation of FSTPSS [29]. The daily operation cost is defined by (12):

\[ C_{\text{DCC}} = \sum_{i=1}^{T} (P_{\text{grid-bus}} \cdot \Delta t / 12 - \sum_{i=1}^{T} (P_{\text{grid-dc}} - \pi_{\text{ac}}) \cdot P_{\text{dc-load}}) \cdot \Delta t / 12 \]  \hspace{1cm} (12)

where \( T \) is the total number of time step in a day, \( t \) is the current time, \( \Delta t \) (5 mins) is the time interval, \( \pi_{\text{ac}} \) (\$/kWh) is the unit price of electricity, \( \pi_{\text{ac}} \) (\$/kWh) is the unit power transmission cost, and \( P_{\text{grid-dc}} \) (kW) represents the active power from the grid consumed by the FSTPSS. The first half of (12) represents the all-day electricity bill for FSTPSS. The second half presents the profit obtained by FSTPSS to supply power to the 10 kV distribution network, and the profit is equal to the price of the current electricity minus the cost of using the transmission line.

Based on the above conditions and optimization goals, if we input the daily forecast data of traction load, PV panel output and 10 kV load, we can get daily optimal reference values for each part of FSTPSS.

D. Objective Function

There are two problems in DAEMS: day-ahead prediction error which is always larger compared to the error of real-time prediction and system transmission loss which is not considered in DAEMS. Therefore, a REMS based on LSTM and DRL is proposed to compensate this error between reference value given by DAEMS and the real value.
It is necessary to have at least one controllable energy storage device that does not participate in the optimization of DAEMS, but specifically participates in REMS. In this paper, UC on the DC side of wayside microgrid is selected as the adjustable device for REMS. This is because the adjustment device needs to be charged and discharged more frequently to compensate the prediction error. Under normal circumstances, the replacement of the UC on the traction station side will cause the traction station to shut down and reduce the power point of TPSS, so the replacement cost will be higher. However, replacing the UC on the DC side of microgrid will not cause this problem. The other adjustable devices are all act according to the optimized value of DAEMS.

The ultimate purpose of the proposed optimization algorithm is to ensure that the powers into and out of the railway follow the reference values set by DAEMS. Therefore, it is reasonable to assume that all other part’s power except 220kV power grid have been operated according to the reference value. Finally, the power error caused by the non-deterministic loads or power sources is eliminated by 220kV power grid. Hence, in the end, if the power of the 220kV grid also meets the reference value, the goal of eliminating the tracking error is achieved.

The schematic diagram of the REMS optimization process is shown in Fig. 2.

REMS uses the LSTM network for real-time prediction, which can reduce the prediction error since the accuracy of real-time prediction is higher than that of day-ahead prediction. The time interval of real-time prediction is 5 seconds. Since the recording reference period of the frequency deviation is 10 seconds, and the calculation period of harmonics and three-phase unbalance is 1 minute according to《Technical guide for the connection of electric railway traction substation to grid》(Q/GDW11623-2017), the selection of 5 seconds time scale can make the final control results meet the requirements. Constraints in FSTPSS and DRL are used to compensate overall error of the system. Among them, we first use the real-time predicted data and the constraints in FSTPSS to derive a rough compensation power \( P_{DC,uc} \), and then, combine with DRL to further derive the correction value of the compensation power \( \Delta P_{DC,ac} \). Concretely, this correction value tries to correct the following error which is represented by (13):

\[
e = L_R(P_{[grid, load, PV, sub, bat, uc, grid, M, load}] - L_d(P_{[grid, load, PV, sub, bat, uc, grid, M, load]} + P_{[sub, bat, uc, grid, M, load]}), (13)
\]

where \( e \) represents the error between the calculated value based on the constraints and the real-time prediction proposed in this paper and the real value. \( P_{[i]} \) indicates the real output of power source or load \( i \) at time \( n+1 \). \( P_{[i]} \) means the predicted output of source or load \( i \) at time \( n+1 \). \( P_{[i]} \) presents the reference output generated by DAEMS. \( P_{DC,uc} \) denotes the expected output of the REMS adjustable device. \( L_d(\cdot) \) represents the real-world functional relationships between the variables and \( P_{sub,grid} \). Theoretically, if \( \Delta P_{DC,uc} \) satisfies (14), it can completely eliminate the error.

\[
\Delta P_{DC,uc} = e
\]

However, before time \( n+1 \), only the values of \( P_{[i]} \) can be determined, and the value of \( P_{[i]} \) at time \( n+1 \) cannot be determined. Thus, it becomes difficult to use (13) directly. Fortunately, the value of \( P_{[i]} \) actually has a great correlation with the value of \( P_{[i]} \). Because \( P_{[i]} \) is the predicted value of \( P_{[i]} \). Intuitively, if the error of the prediction results is small enough, the predicted output variable can actually be the same variable as the real output variable. From a data dimensionality reduction perspective, if two variables are highly linearly dependent, they can be compressed into one variable.

To verify this idea, this paper analyzes the correlation coefficient between these two variables in Appendix and the conclusion is that the variable of the predicted output and the real output can be represented by one variable which could be the variable of the predicted value.

At this point, theoretically, the DRL agent responsible for implementing (14) only needs to complete a function approximator that approximates (13) to achieve unbiased error correction.

Hence, the objective function of this paper can be defined by the following formula:

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} (P_{sub,grid} - L_R(P_{[i]} + P_{DC,uc} + \Delta P_{DC,uc}))^2 \quad (15)
\]

where \( MSE \) means the mean square error (MSE) between the reference value of \( P_{sub,grid} \) and real value of \( P_{sub,grid} \). \( P_{sub,grid} \) presents the reference value of \( P_{sub,grid} \) of sample \( i \). \( P_{[i]} \) denotes
the predicted value of next moment of various non-deterministic loads or sources of sample \( i \). \( P^i_t \) represents the reference value of next moment of other loads or sources of sample \( i \) is the total number of samples.

The above description means that the goal of this paper is to find a suitable REMS to minimize the value described in (15). In this paper, this REMS is implemented based on LSTM and deep Q learning. In subsequent analysis, we will describe in detail how this REMS is implemented.

### III. REAL-TIME ENERGY MANAGEMENT STRATEGY BASED ON LSTM AND DRL

The above description indicates that if a DAEMS is to be implemented reliably, then a REMS is essential which function is to compensate the error of DAEMS. The overall optimization process of REMS consists of two parts: the prediction part and the control part.

LSTM is selected to make predictions in this paper. For the problem in this paper, there are three time series variables that need to be predicted, namely traction load, PV output and 10kV load. Therefore, it can be predicted separately with three LSTM networks. Its time series training samples can be shown in equations (16)-(18):\n
\[
X^i_t = (X^\text{load}_{10kV}^i, X^\text{load}_{DC}^i, X^\text{PV}^i, X^\text{UC}^i, X^\text{grid}^i)\]
\[
P^i_t = (P^\text{uc}^i, P^\text{pod}^i, P^\text{Pv}^i, P^\text{load}^i, P^\text{grid}^i)\]
\[
P^\text{ref}^i = (P^\text{uc}^i, P^\text{pod}^i, P^\text{Pv}^i, P^\text{load}^i, P^\text{grid}^i)\]

where \( X^i_t \) denotes the \( n \)-th time series sample of variable \( i \). \( X^i_t \) presents the value of variable \( i \) at time \( t \). \( N \) is the length of the time series.

The control part of REMS is achieved by RL. RL is an algorithm that learns from interaction with the environment. For the algorithm itself, it has four core elements that need to be defined: state \( s \), action \( a \), reward signal \( R \), and value function \( Q \).

The state variable \( s \) in this paper can be defined by (19-21):

\[
s = (P^\text{uc}_{t+1}^i, P^\text{grid}_{t+1}^i, P^\text{uc}^i)\]

where \( P^\text{ref}^i \) presents the predicted value of each non-deterministic loads or power sources at the next moment. \( P^\text{ref}^i \) denotes the reference value of each controllable power sources or energy storage devices given by DAEMS at the next instant of time, and \( P^\text{ref}^i \) represents the power value of UC at the next instant of time that is calculated from the constraints of FSTPSS.

The action \( a \) which is \( \Delta P^\text{DC,uc} \) in this paper can be defined by (22):

\[
A = \{-K \cdot \Delta P^\text{uc} \cdot (K-1) \cdot \Delta P^\text{uc}, ..., 0, ..., (K-1) \cdot \Delta P^\text{uc}, K \cdot \Delta P^\text{uc} \}
\]

where \( \Delta P^\text{uc} \) is the action value interval which is artificially given. \( K \) is the multiple of the maximum action value relative to the action interval value. Different \( \Delta P^\text{uc} \) and \( K \) will result different performance of RL. The purpose of this action is to eliminate the tracking error described in Section II.

So, the RL agent should get the greatest reward when the tracking error is minimized. Thus, the reward of this article can be defined by (23)

\[
R_{\text{act}} = P^\text{uc,grid} - P^\text{uc,grid}_{t-1}
\]

The RL method used in this paper is an algorithm called Q-learning. Q-learning is a model-free RL method that learns how to find the optimal action selection strategy in a finite Markov Decision Process (MDP). The goal of its agent is to learn an optimal policy from its history of interactions with the environment. The history record is a series of state-action-rewards of the subject [30, 31]. Its value function is to predict the expected return of each state and each action in the future based on the experience of the rewards obtained by the agent’s historical actions. The action-value function \( Q(s,a) \) updating formula of the Q-learning algorithm is as follows:

\[
Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma \max_{A_{t+1}} Q(S_{t+1}, A_t) - Q(S_t, A_t)]
\]

(24)

In (24), \( S_t \) and \( A_t \) respectively represent the state of the environment and the action chosen by the agent at time \( t \). \( \alpha \) presents the learning rate. \( R_{t+1} \) denotes the instant reward obtained by using the action \( A_t \) in the state \( S_t \). \( \gamma \) represents all optional actions in state \( S_{t+1} \). \( \gamma \) is the state of the environment after the action \( A_t \) is selected in the state \( S_t \).

The original Q-learning method is a tabular method, which cannot solve problems such as continuous state problems and dimensional explosions which will cause the memory overflow so that the method will lose its practical meaning. To solve these problems, DQN uses an artificial neural network as a function approximator to approximate the data in the action-value table. The cost function of the Q-network is shown in (25).

\[
L(\theta) = \frac{1}{N} \sum_{i=1}^{N} (R_i + \gamma \max_{\theta} Q(S', A; \theta) - Q(S, A; \theta))^2
\]

(25)

where \( R_i \) are the parameters (or weights) of the Q-networks. \( \theta \) are parameters of the target network which is updated with the Q-network parameters \( (\theta) \) every \( C \) step \( (C \) is a constant value), \( k \) denotes the \( k \)-th sample. \( N \) represents the total number of samples.

The above-mentioned method is called the DQN method using the off-policy strategy. The specific algorithm updating schematic diagram of this DQN method when interacting with the environment is shown in Fig. 3.

Fig. 3. The specific algorithm updating schematic diagram of this DQN.

Specifically, the learning process of deep Q-learning is as follows:

**Step 1**: Obtain the input state of the real world;

**Step 2**: Input the state into the action network to obtain the action;

**Step 3**: Transfer the action to the real world for execution. Afterwards, the corresponding actions, states and rewards are fed back to deep Q-learning;

**Step 4**: Deep Q-learning methods store these actions and rewards in the experience replay pool. The action network will sample in the pool periodically to update its own neural network. In order to solve the stationarity problem when performing the
backpropagation algorithm, the predicted one-step future rewards use an older version of the action network which is target network.

**Step 5**: Determine whether the maximum number of interactions is reached, if yes, end the learning, otherwise return to Step 1.

Furthermore, for action selection strategies with DQN methods, there is a contradiction between exploration and exploitation. That is, we hope that the agent can actively explore while making full use of its experience [32]. The \( \varepsilon \)-greedy policy can well balance the above contradiction, as shown in (26).

\[
\pi(a|s) = \begin{cases} 1 - \varepsilon & \text{if } a = \text{argmax} Q(s,a) \\ \varepsilon / |A(s)| & \text{if } a \neq \text{argmax} Q(s,a) \end{cases}
\]

where \( |A(s)| \) is the number of actions that can be selected in state \( s \).

This policy uses a larger \( \varepsilon \) at the beginning of the iteration. Because there are fewer samples at this time, the policy tends to be actively explored. As the number of iterations increases, \( \varepsilon \) will gradually decay, and the policy for action will be more inclined to the greedy policy.

IV. CASE STUDY

A. Simulation Conditions

In order to validate the effectiveness and superiority of REMS, three simulations were carried out.

The first simulation focuses on the verification of LSTM, by comparing the error between the predicted value and the measured value of non-deterministic loads or sources to verify the effectiveness of LSTM. This is a common performance comparison simulation in the field of load prediction [33].

The second simulation is to compare the algorithm proposed in this paper with some traditional algorithms to verify its effectiveness and superiority. The specific algorithm composition is shown in Table I. This is a set of simulations designed by ourselves after referring to some simulation design experiences of model-free methods [34, 35].

The third simulation tests the influence of the action value selection interval of DQN on its performance.

The parameters of UC and battery involved in DAEMS and REMS are shown in Table II. For the parameters of LSTM, the input of the LSTM network is a sequence of 25 points before the predicted point i.e., the power value within 125 seconds before the predicted time, and the output is the predicted time. The minibatch size is set to 40. The maximum epochs of training are 2000. The initial value of probability of exploration is 1 with a decay rate of 2e-4. The probability will decay every epoch until the value of probability becomes 0.01. The minibatch size is 128. The size of experience replay pool is 50,000. The action value of the DQN is a discrete value of \([-10MW, 10MW]\) in intervals of 1MW. The activation functions of their hidden layers are all the rectified linear unit (ReLU) function. These parameters are designed after being inspired by [34].

The departure interval of trains in the simulation model is shown in Table III. The power drawn by the trains in the feeding section is shown in Fig. 4 as a function of the train distance from the traction substation. The speed of the train is 300km/h. The grid electricity price is assumed constant over the day and equal to ¥0.65/kWh. The electricity transmission fee is ¥0.04/kWh. The equivalent parameters of the traction network after the reduced-order equivalent and decoupling equivalence are (0.13+j0.18) \( \Omega/km \) and (0.23+j0.51) \( \Omega/km \). The former one is the parameters used to implement the model-based method. The latter one is the parameters used to run the simulation. The parameters are calculated from TRANS [36, 37] which is a professional software for TPSS parameters calculation.

![Fig. 4. The relationship between train power and train location in the feeding section.](image)
B. Prediction Results

The prediction results of LSTM network for traction load, PV and 10kV load are shown in Fig. 5.

From Fig. 5, it can be seen that the predicted value of LSTM is very close to the measured value. This shows that the proposed prediction method can predict the random state in this paper very well. In order to further analysis the accuracy of the proposed prediction method, Table IV summarizes MSE and the mean absolute percentage error (MAPE) of the training set and test set of each random state.

<table>
<thead>
<tr>
<th>Prediction object</th>
<th>MSE of training set</th>
<th>MAPE of training set</th>
<th>MSE of test set</th>
<th>MAPE of test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traction load</td>
<td>0.0025</td>
<td>2.41%</td>
<td>0.0026</td>
<td>4.89%</td>
</tr>
<tr>
<td>PV</td>
<td>&lt;0.0001</td>
<td>0.31%</td>
<td>&lt;0.0001</td>
<td>0.32%</td>
</tr>
<tr>
<td>10kV load</td>
<td>0.0001</td>
<td>0.86%</td>
<td>0.0001</td>
<td>0.88%</td>
</tr>
</tbody>
</table>

From Table IV, it can be found that MSE of each random variable is in a very low value. Whether it is the training set or the test set, their value does not exceed 0.01 and this value is acceptable in practice. Moreover, MAPE of each random variable does not exceed 5%. The largest error is MAPE of the test set for traction load, which is close to 5%. The rest errors are all far less than 5%, which is also acceptable in practice. Therefore, the predicted outputs obtained using these LSTM networks can be considered as reliable inputs for control decision part.

C. Control Comparison Results

The results of two comparative simulations are shown in Fig. 6.

D. DQN with Different Action Value Intervals Comparison Results

To further study the influence of the action value of the DQN on its performance, different comparative simulations based on different action value settings have been undertaken.

Therefore, this paper additionally conducts the DQN optimization simulation with the action value interval of 2MW and the action value interval of 0.5MW in the [-10MW,10MW].
The comparison results with the DQN optimization simulation with the action value interval of 1MW are shown below.

![Comparison of tracking errors of various methods](image)

**TABLE V**

<table>
<thead>
<tr>
<th>Action interval</th>
<th>Iteration</th>
<th>Equivalent interaction time</th>
<th>MSE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQN 2MW</td>
<td>2000</td>
<td>33min 20s</td>
<td>0.2205</td>
<td>13.14%</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>1h 23min 20s</td>
<td>0.1938</td>
<td>12.80%</td>
</tr>
<tr>
<td></td>
<td>10000</td>
<td>2h 46min 20s</td>
<td>0.1852</td>
<td>12.58%</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>33min 20s</td>
<td>0.1577</td>
<td>9.52%</td>
</tr>
<tr>
<td>DQN 1MW</td>
<td>5000</td>
<td>1h 23min 20s</td>
<td>0.0904</td>
<td>7.39%</td>
</tr>
<tr>
<td></td>
<td>10000</td>
<td>2h 46min 20s</td>
<td>0.0735</td>
<td>6.99%</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>33min 20s</td>
<td>0.0829</td>
<td>9.52%</td>
</tr>
<tr>
<td>DQN 0.5MW</td>
<td>5000</td>
<td>1h 23min 20s</td>
<td>0.0829</td>
<td>5.47%</td>
</tr>
<tr>
<td></td>
<td>10000</td>
<td>2h 46min 20s</td>
<td>0.0692</td>
<td>4.70%</td>
</tr>
<tr>
<td>Model-based method</td>
<td>--</td>
<td>--</td>
<td>0.1113</td>
<td>12.42%</td>
</tr>
<tr>
<td>No REMS</td>
<td>--</td>
<td>--</td>
<td>9.8832</td>
<td>57.48%</td>
</tr>
</tbody>
</table>

From the results showing above, it can be known that the DQN with smaller action value interval has better tracking performance. The MSE comparison of each case in Table V also proves this conclusion.

It can also be found in Table V that the tracking error of the DQN is related to the number of iterations, that is, the time of interaction with the environment. In general, the longer the interaction time, the better the tracking performance of DQN. However, for DQN with different action intervals, the increasing trend of its tracking performance is not the same. For example, for DQN with 1MW action interval, when the number of iterations is 2000, the tracking error is even smaller than that of DQN with 0.5MW action interval. This is because 1MW has fewer action options, so the performance improvement of its training is faster in the early stage. However, the DQN with 2MW action interval with fewer actions always has a large tracking error. The reason to cause it is that when the action interval is too large, all its actions cannot meet the tracking results more accurately. So even if its performance improves quickly, it cannot have a small tracking error.

Therefore, in the actual selection of DQN action values, if interaction time is not considered, DQNs with smaller action value intervals should be selected. However, if you want to quickly obtain a DQN agent with higher performance, you should choose a DQN with a larger action interval. Nevertheless, the action value interval should be smaller than the allowable error range.

E. Performance Comparison of Each Method Under Different TPSS Parameters

In order to verify the generality of the method proposed in this paper, we used different TPSS parameters to perform simulation and comparison. These parameters are also come from TRANS. The specific reduced-order equivalent and decoupled equivalent traction network parameters are (0.24+j0.61) Ω/km and (0.31+j0.61) Ω/km. The results obtained with the original parameters are labeled Case1, while the results obtained with the two new parameters are labeled Case 2 and Case 3. The results obtained with different simulation parameters are shown in Fig. 8.
inevitably reduced. Therefore, in the actual action value interval the tracking error, but the performance improvement speed is results that the reduction of the action value interval can reduce iteration are compared. It can be seen from the optimization 0.5MW are used respectively, and their tracking error with the selection of the proposed method. In the selection of the action method proposed in this paper is better than the method without REMS, and also simulations. The tracking errors of the power grid using the proposed in this paper is proved through relevant simulation FSTPSS track the reference value set by DAEMS.

Meanwhile, the performance of using real-time energy management strategies with smaller action interval DQN methods is still better than that of using real-time energy management strategy with larger action interval DQN methods. And, the performance of using real-time energy management strategy with smaller action interval DQN methods is still better than that using model-based methods, which verifies the generality of the conclusions of the proposed real-time energy management strategy.

V. CONCLUSION

DAEMS in FSTPSS cannot be reliably achieved in the face of these real-time fluctuation which makes the impact of TPSS on the upper grid uncontrollable. In the area where the power grid is relatively weak, the power is completely obtained from the grid and the use of PQ control will cause the grid voltage in the weak area to fluctuate greatly, which will further lead to control stability problems. To solve these problems, it is necessary to have a REMS that is fine-tuned in real-time based on real-time measured data. Therefore, this paper proposes a REMS, whose goal is to make the real-time operation of FSTPSS track the reference value set by DAEMS.

The effectiveness and superiority of REMS and the method proposed in this paper is proved through relevant simulation simulations. The tracking errors of the power grid using the method proposed in this paper, the model-based method, and the method without REMS are 0.0735, 0.1038 and 9.4177 respectively. And their respectively MAPE are 6.99%, 26.43% and 56.78%. From the simulation results, the method proposed in this paper is better than the method without REMS, and also better than REMS using model-based method.

Meanwhile, this paper also discusses the specific parameter selection of the proposed method. In the selection of the action value interval, the action value interval of 2MW, 1MW and 0.5MW are used respectively, and their tracking error with the iteration are compared. It can be seen from the optimization results that the reduction of the action value interval can reduce the tracking error, but the performance improvement speed is inevitably reduced. Therefore, in the actual action value interval selection, the allowed interaction time should be considered first, and then, on this basis, DQNs with smaller action value intervals should be selected.

In addition, to exclude the effect of TPSS parameters on the results, different TPSS parameters are used to perform simulation and comparison. And the conclusion remains unchanged which verifies the generality of the proposed method.

At present, the strong potential of model-free methods allows us to work on developing a more general control method rather than a model-based one, which has actually caused a complete shift in our research direction, that is, from the original model-driven approach shift to data-driven approach.

APPENDIX

The Analysis of Data Dimensionality Reduction

For multidimensional data, whether it can be dimensionally reduced depends on the correlation between the data. The linear correlation between two variables can be quantified by the correlation coefficient. If the correlation coefficient between them is 1, it means that they are linearly related which means that these two variables could be combined into one variable. Concretely, the formula of the correlation coefficient is as follows:

$$r(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{D}(X)\text{D}(Y)}}$$ \hspace{2cm} (27)

$$\text{Cov}(X,Y) = E(\text{X} \cdot \text{Y}) - E(\text{X})E(\text{Y})$$ \hspace{2cm} (28)

where $X$ and $Y$ are two random arbitrary variables. $r(X,Y)$ is the correlation coefficient between $X$ and $Y$. $\text{Cov}(X,Y)$ presents the covariance between $X$ and $Y$. $E(\cdot)$ is a function that calculate mathematical expectation of random variable.

For the problem of this paper, there are three objects with random characteristics, which are traction load, PV output and 10kV load. Their predicted values are also random variables. We use the same data in Section IV to calculate the actual value and predicted value of each object. The correlation coefficients obtained are shown in Table VI.

**TABLE VI**

<table>
<thead>
<tr>
<th>Object</th>
<th>Correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traction load</td>
<td>0.9956</td>
</tr>
<tr>
<td>PV output</td>
<td>0.9967</td>
</tr>
<tr>
<td>10kV load</td>
<td>0.9745</td>
</tr>
</tbody>
</table>

It can be seen from the table that the correlation coefficients are all very close to 1. In other words, it can be considered that the predicted output variables and real output variables of these three objects are linearly dependent respectively. Therefore, the variable of the predicted output and the real output can be represented by one variable.

REFERENCES

et al.


---

Yichen Ying was born in Zhejiang, China, in 1996. He received the B.S. degree in electrical engineering from University of South China, Hengyang, China, in 2017 and his Ph.D. degree in Electrical Engineering from Beijing Jiaotong University, Beijing, China, in 2023. He was a visiting PhD student in Electrical Power and Control in the Department of Electronic, Electrical, and Systems Engineering from University of Birmingham, Birmingham, U.K, from 2022 to 2023. Now he is currently a Lecturer in Shanghai University of Electric Power. His research interests include power flow calculation, energy management of traction power supply system and AI application in electrified railway.

Zhongbei Tian received the B.Eng in Huazhong University of Science and Technology, Wuhan, China, in 2013. He received the B.Eng. and PhD degree in Electrical and Electronic Engineering from the University of Birmingham, Birmingham, U.K., in 2013 and 2017. He was a Lecturer in Electrical Energy Systems at the University of Liverpool, from 2020 to 2023. He is currently an assistant professor in School of Engineering at the university of Birmingham. His research interests include railway traction power system modelling and analysis, energy-efficient train control, energy system optimisation, and sustainable transport energy systems integration and management.

Mingli Wu was born in Hebei Provence, China, on November 11, 1971. He received his B.S. and M.S. degrees in Electrical Engineering from Southwest Jiaotong University, Chengdu, China, in 1993 and 1996, respectively; and his Ph.D. degree in Electrical Engineering from Beijing Jiaotong University, Beijing, China, in 2006. Since 2008, he has been a Professor in the School of Electrical Engineering, Beijing Jiaotong University. His current research interests include power supplies for electric railways, digital simulation of power systems and electric power quality.

Qiujiang Liu (S’18) was born in Hebei province, China. He received his B.Sc., M.Sc. and Ph.D. degrees in Electrical Engineering from Beijing Jiaotong University (BJTU), Beijing, China, in 2012, 2014 and 2018 respectively. He is currently a Lecturer in BJTU. His research interests include power quality of electric railways, energy storage systems, and other applications of power electronics in traction power supply systems.

Pietro Tricoli (M’06) received the M.S. (cum laude) and Ph.D. degrees in Electrical Engineering from the University of Naples Federico II, Italy, in 2002 and 2005, respectively. He is currently a Professor of Electrical Power and Control in the Department of Electronic, Electrical, and Systems Engineering, University of Birmingham, Birmingham, U.K. His research focusses on modeling and control of power converters with application to storage devices for electric vehicles, railway electrification and traction systems, wind and photovoltaic generation.