Annual Reviews in Control 49 (2020) 145–163

Contents lists available at ScienceDirect

Annual Reviews in Control

journal homepage: www.elsevier.com/locate/arcontrol

Review article

Reinforcement learning in sustainable energy and electric systems: a survey

Ting Yang[a,*], Liyuan Zhao[a,*], Wei Li[b], Albert Y. Zomaya[b]

[a] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
[b] School of Information Technology, The University of Sydney, Camperdown, NSW 2006, Australia

ARTICLE INFO

Article history:
Received 30 December 2019
Revised 15 March 2020
Accepted 18 March 2020
Available online 9 April 2020

Keywords:
Reinforcement learning
Sustainable energy and electric systems
Deep reinforcement learning
Power system
Integrated energy system

ABSTRACT

The dynamic nature of sustainable energy and electric systems can vary significantly along with the environment and load change, and they represent the features of multivariate, high complexity and uncertainty of the nonlinear system. Moreover, the integration of intermittent renewable energy sources and energy consumption behaviours of households introduce more uncertainty into sustainable energy and electric systems. The operation, control and decision-making in such an environment definitely require increasing intelligence and flexibility in the control and optimization to ensure the quality of service of sustainable energy and electric systems. Reinforcement learning is a wide class of optimal control strategies that uses estimating value functions from experience, simulation, or search to learn in highly dynamic, stochastic environment. The interactive context enables reinforcement learning to develop strong learning ability and high adaptability. Reinforcement learning does not require the use of the model of system dynamics, which makes it suitable for sustainable energy and electric systems with complex nonlinearity and uncertainty. The use of reinforcement learning in sustainable energy and electric systems will certainly change the traditional energy utilization mode and bring more intelligence into the system. In this survey, an overview of reinforcement learning, the demand for reinforcement learning in sustainable energy and electric systems, reinforcement learning applications in sustainable energy and electric systems, and future challenges and opportunities will be explicitly addressed.

Contents

* Corresponding authors.
E-mail addresses: yangting@tju.edu.cn (T. Yang), yuaner_zhao@tju.edu.cn (L. Zhao).

https://doi.org/10.1016/j.arcontrol.2020.03.001
1367-5788/

1. Introduction

Energy is an important material basis for the survival and development of human society. In recent years, the global energy demand has been growing strongly, the mismatch between energy supply and demand is becoming increasingly tense, and fossil energy is exhausted day by day. The energy problem has become a severe challenge faced by all countries around the world, which are dealing with the energy crisis and environmental pollution. To solve the common threat faced by all countries in the world, a good number of research has been carried out on the sustainable energy system from two aspects of broadening the sources of energy and reducing the consumption/cost of energy. Broadening the sources of energy is a way of seeking more available energy to maintain the sustainable supply of energy. Besides fossil energy and water energy in the traditional sense, solar energy, wind energy, biomass energy, marine energy, geothermal energy, etc. are highly attractive because of their cleanness and renewability. Reducing the consumption/cost of energy aims to reduce energy waste to the maximum extent, striving to delay the depletion of fossil fuels and reduce environmental pollution by improving the energy utilization rate.

In addition, in the process of social development and progress, energy systems such as oil, gas, electric power and thermal energy are often planned and operated independently for various reasons, and lack of coordination among them. The resulting problems such as low energy utilization rate and low flexibility and reliability of the energy system also need to be solved urgently. At present, the sustainable energy and electric systems are formed by the close coupling of power system, gas system, thermal system and other energy systems, and gradually evolved into an integrated energy system with electric power as the core (Jin, et al., 2016; Zeng, Fang, Li, & Chen, 2016). In this context, several important forms of energy utilization have emerged, such as smart grid (containing microgrid), integrated energy system (IES) and energy internet (EI). Their aims are to build sustainable energy and electric systems through both broadening the sources of energy and reducing the consumption/cost of energy, so as to alleviate the energy crisis and reduce environmental pollution. Power grid is an important part of the energy industry chain and is the hub of energy transmission, especially electric power transmission, which has the functions of optimizing the energy resource allocation mode and improving the energy resource allocation efficiency. The interconnection of power grids and the integration of power, gas, thermal energy and other multi energy sources will form the center of the future sustainable energy system. Renewable energy technology, distributed generation technology, integrated energy utilization technology and energy management technology are developing rapidly in sustainable energy and electric systems. They provide important support for optimizing energy supply structure and improving energy utilization efficiency.

With the integration of intermittent renewable energy and large-scale regional interconnection of energy systems, sustainable energy and electric systems have evolved into highly dimensional dynamic large-scale systems. Moreover, the randomness of users' energy consumption behaviours and the flexible use of active loads (such as electric vehicles) also increase the complexity and uncertainty of energy systems. The sustainable energy and electric systems are in real-time dynamic change, and the research of related problems is nonlinear and uncertain. The primary issue for its research is that it is difficult to establish accurate mathematical models in most cases or to describe it simply by mathematical models (Yang, Zhao, & Wang, 2019).

As an important branch of artificial intelligence and especially machine learning, reinforcement learning (RL) has the advantage of self-learning by interactive trial and error with a dynamic environment and has been used to solve a variety of realistic control decision-making issues in the presence of uncertainty (Rocchetta, Bellani, Compare, Zio, & Patelli, 2019). In intelligent robots, game competitions, industrial manufacturing and other fields, it has shown gratifying application results (Khan, Hermann, Lewis, Pipe, & Melhuish, 2012; Mnih, et al., 2015; Shin, Ryu, & Jung, 2012). RL can learn self-improvement by judging the reward feedback generated by its own experience, so as to learn the optimal policy to achieve the goal (Khan, et al., 2012). It does not require a model of the system dynamics and is not sensitive to the physical model of the study object. Therefore, RL is an extremely valuable tool to make (near-) optimal decisions for nonlinear uncertain systems, in cases when the dynamics are either unknown or affected by significant uncertainty (Buşoniu, de Bruin,

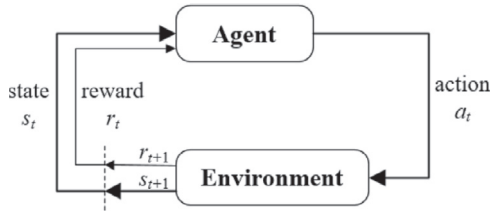

Fig. 1. The interaction between the agent and the environment (adopted from Sutton & Barto, 2018).

Tolić, Kober, & Palunko, 2018). This characteristic of RL makes it very suitable for sustainable energy and electric systems with complex nonlinearity and uncertainty. The applications of RL in sustainable energy and electric systems will effectively improve the economy, flexibility and reliability of the systems and promote the energy revolution.

In a recent review (Glavic, 2019) of (D)RL for electric power system control, the author conducted a comprehensive study on (D)RL for solving power system control and related problems, which offers a good reference for researchers interested in this field. It focuses on power system control, but system decisions making (like energy scheduling and market decisions) are not included. Differently, our paper reviews the applications of (D)RL in energy and electric systems, and provides a comprehensive study on (D)RL for decision-making and control problems. Towards building sustainable energy and electric systems, how to reduce the consumption/cost of energy is also an important aspect. Therefore, for different stakeholders, we give a detailed review of the system operation optimization (i.e. energy management of the energy supply side, the user side and the whole system). In addition, the applications of RL in electric vehicle management is also introduced.

This paper describes RL, the supporting role of RL in sustainable energy and electric systems and the applications of RL in sustainable energy and electric systems, in order to provide useful reference for the development of intelligent energy. The remainder of this paper is organized as follows. Section 2 describes the theoretical basis of classical RL. Section 3 introduces deep RL (DRL) and some popular DRL algorithms. The supporting role of RL in sustainable energy and electric systems is discussed in Section 4. Then, Section 5 summarizes the applications of RL in sustainable energy and electric systems from six aspects. Section 6 describes the challenges and prospects of the applications.

2. Classical reinforcement learning

The core of RL is to learn the dynamic interaction between the agent and the environment where the agent operates. If an agent's action causes the environment to give a positive reward, the agent's subsequent tendency to execute this action will be reinforced. As shown in Fig. 1, the interaction process between the agent and environment can be represented by the closed loop (Sutton & Barto, 2018). The paradigm of RL is very similar to the process of human learning knowledge, which is why RL is regarded as an important way to realize general artificial intelligence. The goal of RL is to find the optimal policy in each state to maximize the long-term reward (discounted). Therefore, RL can learn self-improvement only by judging the feedback from its own experience in the environment, which has a stronger self-learning ability than other machine learning methods.

2.1. Markov decision process

The classical research of RL is based on Markov decision processes (MDPs) (Sutton & Barto, 2018), that is, the next state of the system only depends on the current state and action. An MDP is defined as a tuple $<S, A, T, r>$, where: S is the state space, A is the action space, $T: S \times A \times S \rightarrow [0,1]$ is the state transition probability function which represents the probability of transiting from one state to another after executing an action in that state, $r: S \times A \rightarrow r$ is the reward function which represents the immediate reward received by the agent after it performs the transition.

The agent observes the current state $s \in S$ and chooses an action $a \in A(s)$, where $A(s)$ is the set of all admissible actions in state s. The system evolves probabilistically to the next state $s' \in S$ according to the transition probability function $T(s, a, s')$. The agent then receives reward $r(s, a)$ for the transition. This process can go on indefinitely or end at the terminate state. The policy π is a mapping from a state to an action: $S \times A \rightarrow [0,1]$. The agent's goal is to maximize its (discounted) cumulative reward over the time by adjusting the policy. Generally, two kinds of value functions can be used to measure the quality of a policy π, that is, the state value function $V^{\pi}(s)$ and the state-action value function $Q^{\pi}(s, a)$. Given a policy π, the state value function is defined as follows:

$$V^{\pi}(s) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s\right] \tag{1}$$

where $E_{\pi}[\cdot]$ is the expectation under policy π, s_0 is the initial state, t is the index of time step, r_t is the reward at time step t, and the parameter $\gamma \in [0,1]$ is the discount factor used to reduce the effect of future reward while keeping the cumulative reward over infinite horizon bounded.

The state-action value function is defined as follows:

$$Q^{\pi}(s, a) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a\right] \tag{2}$$

where a_0 is the initial action. Since the optimal policy π^* is the policy which maximizes the value function, the optimal policy of each value function can be obtained according to the following two equations:

$$\pi^* = \arg\max_{\pi} V^{\pi}(s) \tag{3}$$

$$\pi^* = \arg\max_{a \in A} Q^{\pi}(s, a) \tag{4}$$

2.2. Temporal difference

In Section 2.1, the MDP is defined as a tuple $<S, A, T, r>$. According to whether the state transition probability T is known or not, RL methods can be divided into model-based RL methods and model-free RL methods. Model-based methods rely on planning as their primary component, while model-free methods primarily rely on learning. The model-based RL can be solved by dynamic programming (DP) while model-free RL can be solved by Monte Carlo (MC) and temporal difference (TD) methods (Sutton & Barto, 2018). The implementation of DP needs a complete environment model; while MC method can be used without a model, but it uses experience to evaluate the value function and must wait until the end of the episode to exploit the return, so the learning speed is slow and it is difficult to realize online. The TD method, which combines DP and MC methods, is a model-free RL method and can be realized online through step-by-step incremental computation (Sutton, 1988).

TD method is one of the most popular RL schemes. Among TD methods, the most classic is Q-learning proposed by Watkins in his doctoral thesis in 1989 (Watkins, 1989). The main idea of the algorithm is to define Q-functions and substitute the data observed online into the following updating formula to learn the Q-functions iteratively and get the exact solution:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t \delta_t \tag{5}$$

$$\delta_t = r_{t+1} + \gamma \max_{a'} Q_t\left(s_{t+1}, a'\right) - Q_t(s_t, a_t) \tag{6}$$

where a_t is the learning rate, δ_t is the TD error, a' is the action that can be taken in state s_{t+1}. Q-learning is an off-policy learning algorithm.

Another RL algorithm similar to Q-learning is SARSA (State-Action-Reward-State-Action) proposed by Rummery and Niranjan (1994). Different from Q-learning, SARSA is an on-policy TD method which updates Q-function directly with online actions. Its TD error is defined as

$$\delta_t = r_{t+1} + \gamma Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t) \tag{7}$$

It can be seen from Eq. (7) that SARSA needs to use the five parts (s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1}) when updating Q-function, which constitute the name SARSA of the algorithm. Under certain conditions, SARSA can obtain the optimal control policy when time tends to infinity (S. Singh, Jaakkola, Littman, & Szepesvári, 2000).

Both Q-learning and SARSA use TD error to update the value function, which belong to TD learning (Sutton, 1988). This involves temporal credit assignment, that is, how much TD error should be assigned to the actions at different times to update the value function. So Sutton (1988) proposed TD(λ) algorithm to solve the problem of temporal credit assignment:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha \delta_t e_t(s_t) \tag{8}$$

where $e_t(s)$ is the eligibility trace, defined as

$$e_t(s) = \begin{cases} \gamma \lambda e_t(s) + 1, & s = s_t \\ \gamma \lambda e_t(s), & \text{else} \end{cases} \tag{9}$$

TD (λ) establishes a unified framework of MC and TD learning. When λ takes different values from 0 to 1, TD(λ) can be converted into different methods. When λ is 1, TD(λ) is MC, while when λ is 0, TD(λ) is TD learning.

2.3. Policy gradient learning

The RL methods introduced above are all value-function-based algorithms, which need to find out the value function first and then improve the current policy based on the value function. Another kind is policy-based methods, such as policy gradient algorithms (Degris, Pilarski, & Sutton, 2012; Williams, 1992). Policy gradient algorithms are methods using approximators to approximate and optimize policy so as to find the optimal policy. These methods can be divided into deterministic policy gradient (DPG) algorithm and stochastic policy gradient (SPG) algorithm. In DPG algorithm, action is executed with probability 1, while in SPG algorithm, action is executed with a certain probability.

Silver, et al. (2014) proposed an effective DPG estimation method. Compared with SPG, the DPG has better performance in high-dimensional action space. Suppose the policy that needs to be approximated is $\pi(s, a; \theta)$ and the policy is derivable with parameter θ. Then the objective function and value function can be defined as

$$J(\pi_\theta) = \mathrm{E}\sum_{t=1}^{\infty} \gamma^{t-1} r_t | s_0, \pi_\theta \tag{10}$$

$$Q^{\pi_\theta}(s, a) = \mathrm{E}\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} | s_t = s, a_t = a, \pi_\theta \tag{11}$$

Suppose that starting from the initial state s_0, the state distribution of the action is selected according to the policy π_θ as

$$d^{\pi_\theta}(s) = \sum_{t=1}^{\infty} \gamma^t P(s_t = s | s_0, \pi_\theta) \tag{12}$$

Then we can derive the following policy gradient theorem(Sutton, McAllester, Singh, & Mansour, 2000):

$$\nabla_\theta J(\pi_\theta) = \sum_s d^{\pi_\theta}(s) \sum_a \nabla_\theta \pi_\theta(s, a) Q^{\pi_\theta}(s, a) \tag{13}$$

From the above formula, it can be seen that although the state distribution $d(s)$ is related to the policy π_θ, $\bigtriangledown\theta J(\pi\theta)$ has no relation with$\nabla_\theta d^{\pi_\theta}(s)$. Therefore, the change of policy will change the sample distribution, but the change of sample distribution will not affect the update of the policy. After obtaining the policy gradient, methods such as gradient ascent can be used to maximize the objective function.

The previous contents introduce value-function-based RL methods and policy-based RL methods. Barto, Sutton, and Anderson (1983) combined the advantages of value-function-based methods and policy-based methods for the first time and proposed Actor-Critic RL. Actor-Critic is composed of an actor network (policy) and a critic network (value function). The actor network defines parameterized policies and outputs actions to the environment based on the observed state of the environment and the critic network outputs the value function based on the state and the return from the environment. The TD error generated by critic network drives the operation of the whole network.

2.4. Multi-agent reinforcement learning

In the previous introduction, we mainly discuss the RL methods for single-agent cases. In practice, in a variety of domains, there are situations where multiple agents need to work together, e.g. robotic teams, distributed control, cooperative driving of multiple vehicles and resource management. In these cases, RL methods for multi-agent systems are designed. For example, in resource management, while resources can be managed by a centralized approach, defining each resource with an agent and using a distributed control strategy may provide a helpful and distributed perspective on the system.

A multi-agent system consists of a group of autonomous, interacting agents sharing a common environment, and has a good degree of scalability (Bu, Babu, & De Schutter, 2008). The multiple agents in the system can interact with each other in perfectly competitive, hybrid strategic, or completely cooperative settings.

3. Deep reinforcement learning

Although classical RL methods have achieved great success in fields of natural science and engineering, many methods are aimed at discrete states and actions. However, in many practical applications, the quantity of states (and actions) is large, even continuous state (and action) spaces. At this time, classical RL methods are difficult to learn effectively. As a result, the classical RL methods suffer from the “curse of dimensionality” problem. Therefore, researchers try to overcome this problem from different perspectives, among which function approximation is the most direct solution by using a function to approximate the value function and/or the policy.

In recent years, deep learning (DL), which has made great progress, has been used to approximate arbitrary complex functions. Due to the excellent properties of adaptability, advanced input-output mapping and nonlinearity, DL is frequently used for universal function approximation in numerical algorithms. The combination of DL function approximators with RL into DRL is tempting, especially for domains such as resource management and robotics where it can enable learning behaviors directly from raw states through trial and error and achieves end-to-end learning. Mnih, et al. (2013) introduced a new DL model for Q-learning, and demonstrated its ability to master difficult control policies

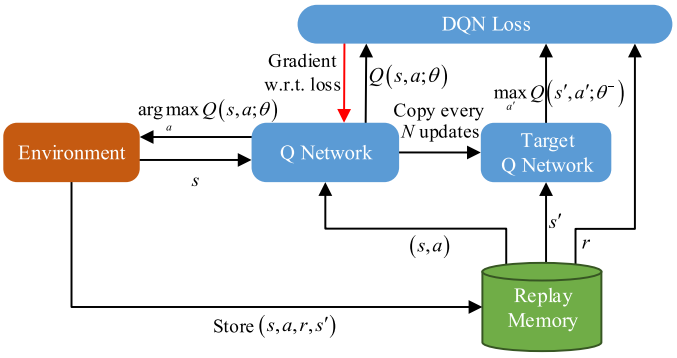

Fig. 2. The schematic diagram of the DQN training process (adopted from Liu, Gao & Luo, (Liu et al., 2019)).

for Atari 2600 computer games. In 2019, University of California, Berkeley and Google Brain enabled quadruped robots to learn to walk like human beings through DRL. The above research shows that DRL has good performance in dealing with decision-making problems with continuous state space.

3.1. Deep learning

The concept of DL originates from the research and development of artificial neural networks (ANNs). A multilayer perceptron with multiple hidden layers is a typical DL structure, where the first several hidden layers can automatically construct new features from the data in an unsupervised way, and then extract more abstract high-level category attributes layer by layer so as to discover the deep feature representation of the data.

Similar to traditional machine learning methods, DL can be divided into supervised learning and unsupervised learning (LeCun, Bengio, & Hinton, 2015). For instance, convolutional neural network (CNN) (Hu, Lu, Li, & Chen, 2014) is a kind of machine learning model under deep supervised learning, while deep belief net (DBN) (Hinton, Osindero, & Teh, 2006), stacked auto-encoder (SAE) (Jiao, Huang, Ma, Han, & Tian, 2018) and restricted Boltzmann machine (RBM) (Hinton, 2012) are machine learning models under unsupervised learning.

DRL which uses DL to approximate the value function is called value-function-based DRL while DRL which uses DL to approximate the policy and solved by policy gradient methods is called policy-based DRL (Arulkumaran, Deisenroth, Brundage, & Bharath, 2017). We will summarize them separately in Section 3.2 and Section 3.3.

3.2. Value-function-based DRL

Q-learning is a classical RL algorithm of learning value function. Next, we will focus on Deep Q-network (DQN), the originator of DRL, and various extension methods of DQN.

3.2.1. Deep Q-network

Deepmind proposed the DQN model in 2013 (Mnih, et al., 2013), a pioneering work in DRL, which trained CNN with a variant of Q learning. Then they added a target network to improve the learning performance of DQN (Mnih, et al., 2015). Before the appearance of DQN, when using neural networks to approximate the value function in RL, there will be problems of instability or even non-convergence. To solve these problems, DQN uses two technologies: experience replay mechanism and target network. The schematic diagram of the DQN training process is shown in Fig. 2.

Experience replay mechanism: In supervised learning with great progress in DL, all the samples are independently and identically distributed. However, the samples in RL are highly correlated and non-stationary, resulting in slow training convergence. The experience replay mechanism stores the agent's experience $e_t = (s_t, a_t, r_t, s_{t+1})$ at each time step and forms a replay memory sequence $D=\{e_1, \ldots, e_N\}$. During the training, a small batch of experience samples is randomly extracted from D each time, and the network parameters are updated by using the stochastic gradient descent algorithm. The experience replay mechanism breaks the correlations between the samples by randomly sampling historical data, while the reuse of experiences increases the efficiency of data usage.

Target Network: In order to make the performance of the DQN algorithm more stable, two neural networks with the same structure are established: the network Q-network which keeps updating the parameters of the neural network and the target Q-network which is used to update the Q-value. DQN uses deep neural network (Q-network) with parameters θ to estimate the action-value function, expressed as $Q(s,a;\theta) \approx Q^{\pi}(s,a)$. During the training process of the Q-network, the parameters are updated by minimizing the following loss functions:

$$L_t(\theta_t) = \mathrm{E}_{(s,a,r,s')}\left[(y - Q(s,a;\theta_t))^2\right] \tag{14}$$

where the optimization target value y of Q-network is:

$$y = r + \gamma \max_{a'} Q\left(s', a'; \theta_t^-\right) \tag{15}$$

where $\theta-$ tare the parameters of the target network. Differentiating the loss function with respect to the weights we arrive at the following gradient:

$$\nabla_{\theta_t} L_t(\theta_t) = \mathrm{E}_{(s,a,r,s')}\left[(y - Q(s,a;\theta_t))\nabla_{\theta_t} Q(s,a;\theta_t)\right] \tag{16}$$

The structure of the target Q-network is the same as that of the Q-network, while the parameters $\theta-$ tof the target Q-network are updated with the parameters θ_t every N iterations and are kept fixed in each time period.

3.2.2. Extension algorithms of DQN

Since DQN was proposed, various extended versions of DQN have emerged. According to the different emphasis of the extensions, the extensions mainly include the improvement of the training algorithm, the improvement of the neural network structure and the improvement of the learning mechanism. Following we will introduce typical algorithms of these extensions.

3.2.3. Double DQN (DDQN)

In Q-learning and DQN algorithms, the parameters are updated as

$$\theta_{t+1} = \theta_t + \alpha\left(y_t^Q - Q(s_t, a_t; \theta_t)\right)\nabla_{\theta_t} Q(s_t, a_t; \theta_t) \tag{17}$$

where the optimization target value is:

$$y_t^Q = r_{t+1} + \gamma Q\left(s_{t+1}, \arg\max_a Q(s_{t+1}, a; \theta_t); \theta_t\right) \tag{18}$$

From Eq. (18), it can be seen that DQN uses the same parameters θ_t to select and evaluate an action, which makes it more likely to overestimate the Q-value.

DDQN (van Hasselt, Guez, & Silver, 2016) uses double Q-learning (van Hasselt, 2010) for reference to modify DQN algorithm. It uses the online Q-network to select an action and uses the target network to estimate the Q-value, which makes the optimization target value:

$$y_t^{\mathrm{DDQN}} = r_{t+1} + \gamma Q\left(s_{t+1}, \arg\max_a Q(s_{t+1}, a; \theta_t); \theta_t^-\right) \tag{19}$$

where θ_t are the parameters of the online network, $\theta-$ tare the parameters of the target network.

3.2.4. Dueling DQN

The traditional DQN architecture consists of input layer, convolution layers (or recurrent layers) (Glavic, 2019), fully connected layers and output layer. Z. Wang, Schaul, et al. (2016) proposed a new neural network architecture for mode-free RL: dueling architecture. It divides the abstract features extracted from convolution layers into two streams in the fully connected layer: one stream represents the scalar state value function $V(s)$, and the other stream represents the action advantage function $A(s,a)$ in a certain state. Finally, the two streams are combined via a special aggregating layer to produce an estimate of the state-action value function. The advantage of this design is to generalize learning across actions without imposing any change to the underlying RL algorithm.

In Dueling DQN, the constructed Q-function is as follows:

$$Q(s,a;\theta,\alpha,\beta) = V(s;\theta,\beta) + A(s,a;\theta,\alpha) \tag{20}$$

where α and β are the parameters of the two streams respectively, and θ are the parameters of DQN.

3.2.5. Prioritized replay DQN

DQN and DDQN mentioned above both remember and reuse past experiences through an experience replay mechanism, and sample transitions from the replay memory. However, these methods use the same frequency to replay the previously experienced transition (s_t, a_t, r_t, s_{t+1}), and do not consider the importance of each transition.

Based on this, Schaul, Quan, Antonoglou, and Silver (2016) proposed prioritized experience replay to replay important transitions more frequently, so that the agent can learn more effectively from some transitions than from others. The importance of experience transitions is measured by TD error. The samples with larger absolute value of TD error are more likely to be sampled for training.

3.3. Policy-based DRL

Policy-based DRL methods mainly include deep DPG (DDPG), trust region policy optimization (TRPO), asynchronous advantage Actor-Critic (A3C), and so on.

3.3.1. DDPG

DQN can only deal with tasks with discrete and low-dimensional action spaces, but in many cases, especially physical control tasks, the action spaces are continuous and high-dimensional. An effective way to apply DRL methods such as DQN to a continuous domain is to discretize the action space, but it will bring a significant problem of curse of dimensionality: the number of actions increases exponentially with the increase of degree of freedom, which will bring great difficulties to training. In addition, simply discretizing the action space will unnecessarily remove information of the action domain. Lillicrap, et al. (2015) applied the idea of DQN to a continuous action domain, and proposed a model-free algorithm DDPG based on DPG and Actor-Critic.

DDPG uses two independent networks to approximate critic (value) function (θ^Q) and actor (policy) function (θ^π), and each network has its own target network $\theta^{Q'}$ and $\theta^{\pi'}$. $\theta^{Q'}$ and $\theta^{\pi'}$ are updated in soft-update techniques to further increase the stability of the learning process:

$$\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'} \tag{21}$$

$$\theta^{\pi'} \leftarrow \tau\theta^{\pi} + (1-\tau)\theta^{\pi'} \tag{22}$$

3.3.2. TRPO

While DDPG is an off-policy DRL technique, Schulman, Levine, Abbeel, Jordan, and Moritz (2015) put forward a policy optimization algorithm, trust region policy optimization (TRPO), with guaranteed monotonic improvement. The problem TRPO trying to solve is to choose the appropriate step size by introducing the trust region constraint defined by Kullback-Leibler divergence, so as to ensure that the optimization policy produces a result never getting worse than the previous ones. To perform the optimization, a linear approximation is made to the objective and a quadratic approximation is made to the constraint. Then the conjugate gradient algorithm is used to calculate the next parameter vector followed by a line search.

3.3.3. A3C

Mnih, et al. (2016) proposed a lightweight framework for DRL that used asynchronous gradient descent for optimization of DL controllers, which used parallel actor-learners to update a global shared model. The framework is combined with four standard RL algorithms (one-step SARSA, one-step Q-learning, n-step Q-learning, and advantage Actor-Critic) to form their own asynchronous variants. First, it uses multiple CPU threads on a single machine, which removes the communication costs of sending gradients and parameters and enables use of a lock-free style (Recht, Re, Wright, & Niu, 2011) for training. Then parallel multiple actor-learners are adopted to explore different parts of the environment, and different exploration policies can be adopted in each actor-learner to maximize the diversity. By running different exploration policies in different threads, different actor-learners update online parallel parameters, which reduces the correlation of data in time. Therefore, a stable learning process can be achieved without an experience replay mechanism. A3C operates in the forward view and uses the same mix of n-step returns to update both the policy and the value-function. The policy and the value function are updated after every t_{max} actions or when the terminal state is reached.

In addition, popular policy-based DRL methods also include proximal policy optimization (PPO) (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017), Actor Critic with experience replay (ACER) (Z. Wang, Bapst, et al., 2016) and so on, which are not covered in this section considering the length of the paper.

4. The supporting role of RL in sustainable energy and electric systems

In the process of the evolution of a power system, which is the core part of sustainable energy and electric systems, to smart grid, EI and IES, it has become a highly-dimensional system with complex structure, various equipment and complex technologies, which has complex nonlinearity and uncertainty. In addition, the further integration of renewable energy makes sustainable energy and electric systems increasingly complex. There are two main reasons: on the one hand, the output of renewable energy has strong spatio-temporal uncertainty, which makes the system operation far more complex; on the other hand, the grid connection, transmission and accommodation of renewable energy introduce more power electronic devices into the system, which will cause the reduction of system inertia and change of its stabilization mechanism. In respect of energy consumption, the emerging loads are increasingly implicated in the development of the load. For example, the short-term peak power consumption caused by the spatio-temporal uncertainty of electric vehicle charging demand is extremely prominent. In sustainable energy and electric systems, the various forms of energy production, the interactivity and uncertainty of energy consumption patterns are integrated,

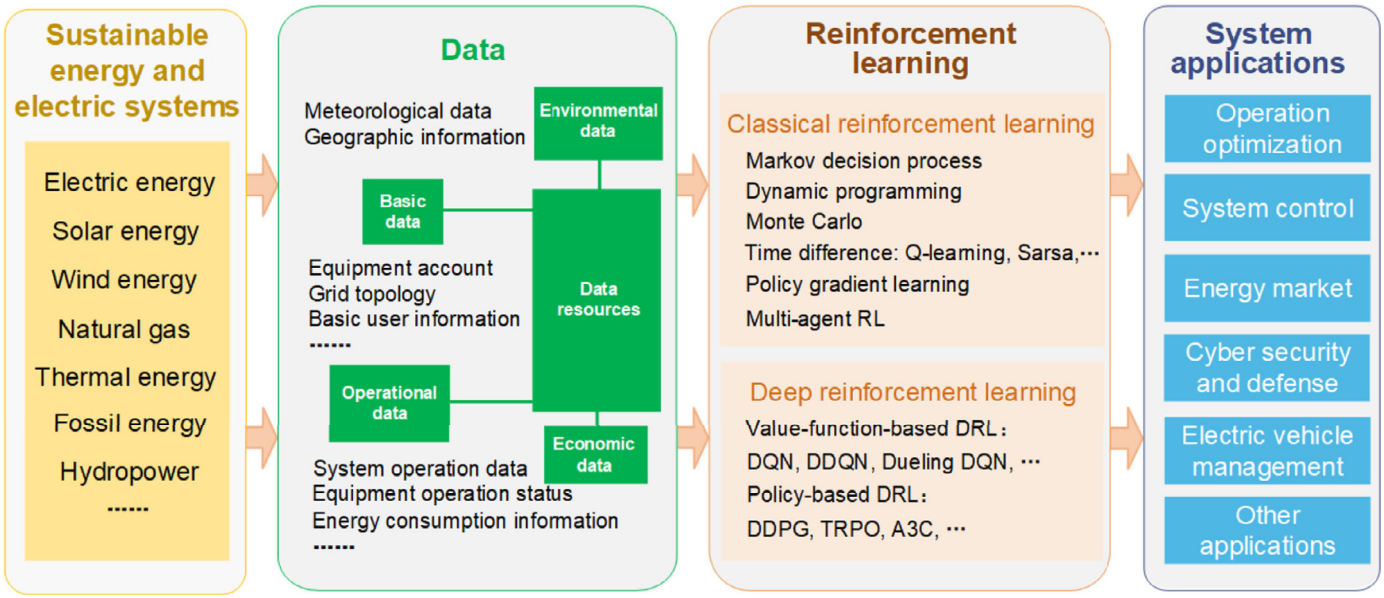

Fig. 3. The supporting role of RL in sustainable energy and electric systems.

which make the system become a multi-objective function constrained by balance. The factors of the system are numerous, the conditions are stochastic, and it is difficult to physically model. The traditional methods of physical modeling or engineering experience have become inadequate. There are still many scientific challenges and technical problems in building efficient sustainable energy and electric systems and supporting multi-energy coordination, optimization and complementarity. RL does not need the model of system dynamics. It can transform the performance index of the system into an evaluation index and learns through reward feedback, which will effectively solve various challenges faced by sustainable energy and electric systems.

Besides, in order to ensure the stable, economic and reliable operation and optimal management of the system, more and more intelligent electronic devices are connected to the energy system, forming a large amount of data resources with wide types and high volumes. Big data drives the sustainable energy and electric systems into a digital era. Massive data provides available environmental state information for the application of RL in sustainable energy and electric systems. These are expected to benefit the "end-to-end" energy management in sustainable energy and electric systems, and also help users transition from a passive to an active role.

Therefore, RL is an important strategic support for energy transformation and development and building sustainable energy and electric systems. Under the pattern of profound evolution of energy structure, the deep integration of RL and sustainable energy and electric systems will effectively improve the ability to control complex energy systems and improve the reliability and economy of system operation. Fig. 3 shows the relationship among sustainable energy and electric systems, system data resources, RL and the applications of RL in various aspects in the system. Based on the environmental data, operational data (for example: energy consumption information), basic data (for example: user information), and economic data in sustainable energy and electric systems, various system applications can be realized with the help of RL technologies such as classical RL methods and DRL methods.

5. Applications of RL in sustainable energy and electric systems

5.1. Applications of RL in operation optimization of sustainable energy and electric systems

Integration of renewable energy generation units and new power electronic devices, randomness of users' energy consumption behaviors and the increase of active loads represented by electric vehicles inject more uncertainty into sustainable energy and electric systems, which makes the optimization of system operation more complex and difficult.

1. Due to the access by various new devices, the network structure of the energy system and the operation mode of each part are flexible and changeable, which means its operation problems have higher requirements for flexibility.
2. The interaction between the demand side and the power/energy grid significantly enhances the system uncertainties, which adds new difficulties to the optimal operation of the system.
3. The investors in sustainable energy and electric systems will be diversified in the future, which may be the government or users themselves, or the independent energy service provider and their combination. The uncertainty of investors will lead to more complexity and changeability in the operation mode, which makes it difficult to accurately consider the operation economy of the system.
4. In order to optimize the system operation, besides the traditional economy and reliability objectives, the optimal operation of sustainable energy and electric systems should also consider other operational objectives such as improving the efficiency of comprehensive energy utilization, minimizing the environmental pollution and maximizing social benefits. Many of these factors are related to social, economic, policy and human constraints which are difficult to quantify. Fig. 4 shows the coupling of key factors to be considered when formulating optimal operation strategies for sustainable energy and electric systems.

Therefore, the operation optimization of sustainable energy and electric systems forms a non-linear and uncertain optimization problem with multi-dimensional variables, multi-objectives and

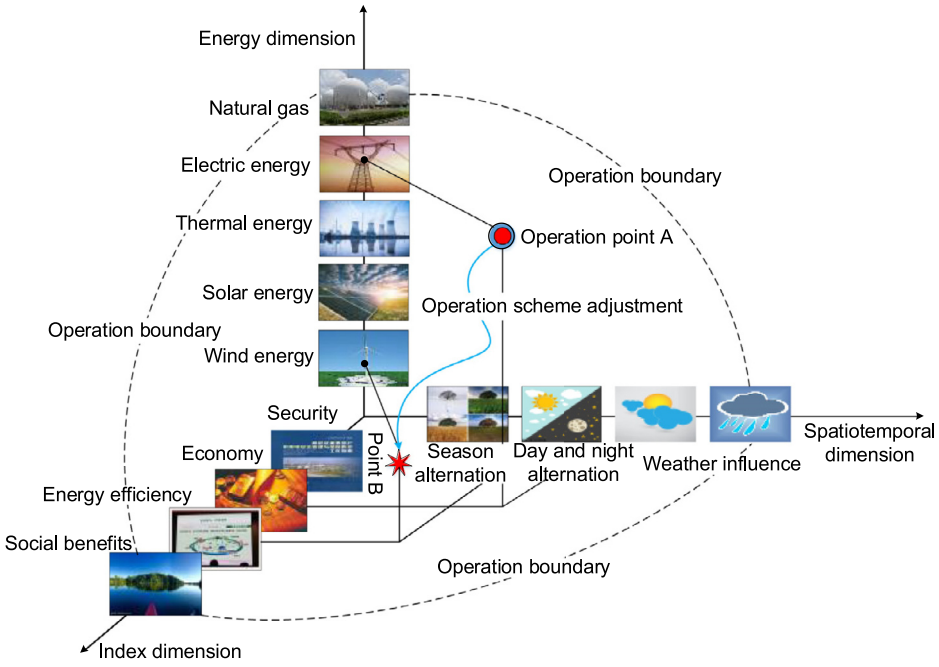

Fig. 4. Key factors to be considered in operation optimization of sustainable energy and electric systems.

Table 1
Summary of applications of RL in operation optimization of sustainable energy and electric systems.

Classification	Energy system/ device	Learning algorithm	Reference(s)
Energy supply side	Single microgrid	Fuzzy Q-learning (multi-agent)	Kofinas et al. (2018)
		Cooperative RL (multi-agent)	Weirong Liu, et al. (2018)
		DQN	Ji et al. (2019)
		Combination of ADP and RL	Venayagamoorthy et al. (2016)
	Multi-microgrid	Combination of DNN and MC	Du & Li (2020)
	EI	A3C	Hua et al. (2019)
	IES	Q-learning	Y. Li et al. (2018), H. Liu et al. (2019), Rayati et al. (2015)
	Single ESS	DDQN	Bui et al. (2020)
	Hybrid ESSs	Q-learning	X. Qiu et al. (2016)
		DDQN	Z. Zhang et al. (2019)
User side	Electric water heater	Q-learning	Al-Jabery et al. (2014), Al-Jabery et al. (2017)
		Fitted Q-iteration	Ruelens et al. (2018)
	HVAC	Q-learning	Sun et al. (2015), Yujiao Chen et al. (2018)
		Combination of RNN and Actor-Critic	Y. Wang et al. (2017)
	Battery	Dual iterative Q-learning	Wei et al. (2015)
	Residential customers in IES	Q-learning	Sheikhi et al. (2016)
	Thermostatically controlled loads	Batch RL (multi-agent)	Claessens et al. (2018)
	Smart building(s)	DQN/ Deep Policy Gradient	Mocanu et al. (2019)
	Different home appliances	Q-learning (multi-agent)	Lu et al. (2019)
Whole system	Wind resources, PV, diesel generators, battery, and customers	Q-learning (multi-agent)	Foruzan et al. (2018)

multi-constraints, which can be solved by RL. The goal of RL is to maximize the long-term average reward of a generic system. This coincides with the design objective of the operation optimization which is established to improve the operation performance of sustainable energy and electric systems at a setting stage. This section divides the operation optimization of the energy systems into three parts according to the research object: operation optimization of energy supply side, operation optimization of the user side and operation optimization of the whole system. Table 1 summarizes these applications.

5.1.1. Operation optimization of the energy supply side

Under uncertainties introduced by intermittent renewable energy and stochastic demand from the consumers, RL can help sustainable energy and electric systems cope with the challenges of guaranteeing energy supply and increasing the reliability of the system. For instance, in order to solve the complex operation optimization problem on the energy supply side of a stand-alone microgrid, each component of the microgrid was deployed as an independent agent to reduce the state space in (Kofinas, Dounis, & Vouros, 2018). The multi-agent system adopts a distributed and collaborative RL method to learn the optimal strategy. Furthermore, in order to control the microgrid components which operate in a continuous states and action space, fuzzy Q-learning is introduced for individual agents to learn to manage system components. In (Venayagamoorthy, Sharma, Gautam, & Ahmadi, 2016), an optimal or near-optimal intelligent dynamic energy management system (I-DEMS) which could perform grid-connected and islanded microgrid operations was developed by using an evolutionary adaptive DP (ADP) and RL framework. The framework includes two neural networks, among which the action network uses the microgrid's states to generate energy dispatch control signals, while the critical network evaluates the dispatch signals over time, realizing fast online dynamic optimization of the I-DEMS performance. In addition,

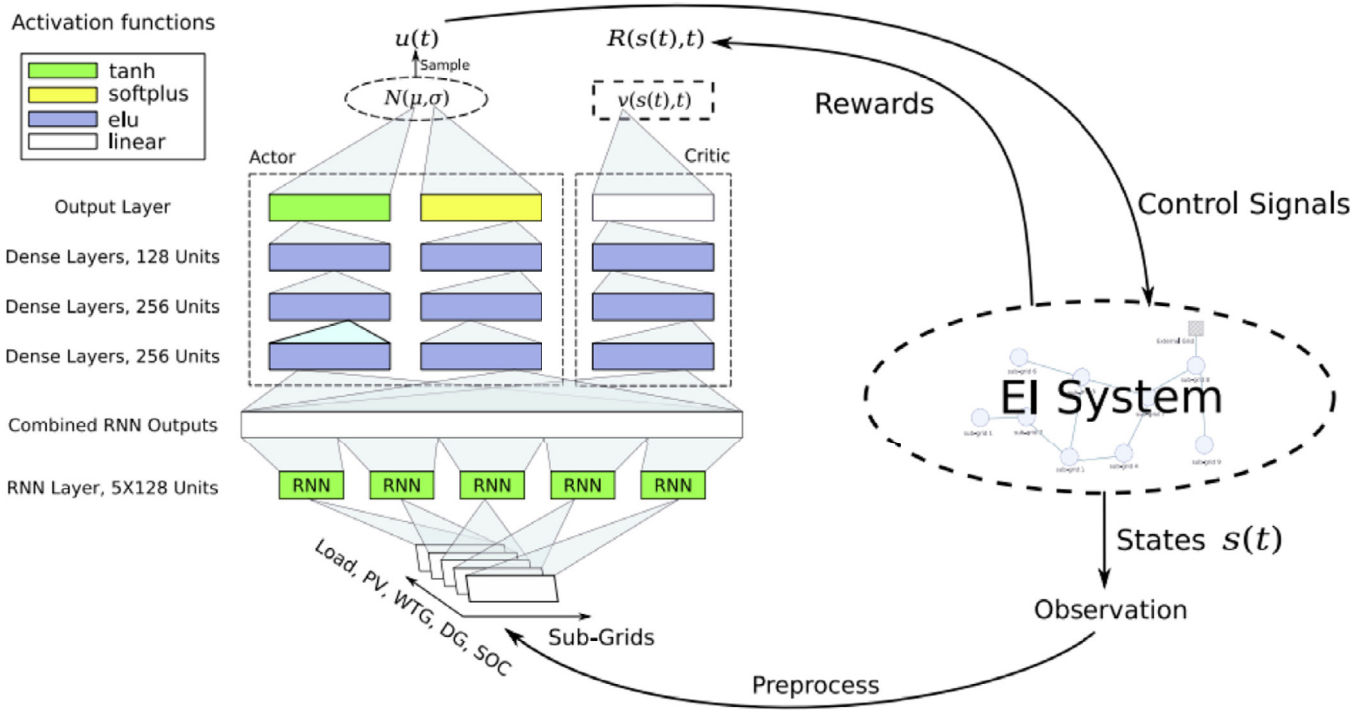

Fig. 5. Structure of EI operation optimization based on A3C in (Hua, Qin, Hao, & Cao, 2019).

Weirong Liu, Zhuang, Liang, Peng, and Huang (2018) and Ji, Wang, Xu, Fang, and Zhang (2019) also conducted in-depth research on operation optimization of the energy supply side of a microgrid based on cooperative RL and DQN respectively.

The above research aims at the operation optimization of single microgrid, for the operation optimization of multi-microgrid (MMG), Du and Li (2019) considered customer privacy and proposed an intelligent MMG energy management method by implementing deep neural network (DNN) and model-free RL techniques without direct access to user's information, which could solve the problem of power system operation optimization with hidden information or large search space. With respect to EI operation optimization, considering multi-objectives and multi-constraints, the EI operation optimization problem can be formulated as a constrained optimal control problem. Since it is difficult to establish accurate mathematical models for the power of renewable power resources and load, A3C algorithm is used to solve the optimal control problem of EI (Hua, Qin, Hao, & Cao, 2019). By applying the A3C algorithm designed in Fig. 5, the intelligent controller for the EI can be reached. Given an observation of the EI system, the network will generate corresponding control signals to achieve an intelligent operation. The states of the environments include load, photovoltaic power generation (PV), wind turbine power generation (WTG), distributed generation (DG) and state of charge (SOC) of the battery energy storage (BES); the actions include control signals of micro-turbine generator (MTG), diesel engine generator (DEG), fuel cell (FC), and BES, and the optimization goal is to find the optimal control scheme such that the sum of total operation cost and penalty cost is minimized, for more detailed information, please see (Hua, et al., 2019).

In addition, for the operation optimization of the energy supply side of IESs, (Rayati, Sheikhi, & Ranjbar, 2015) adopted RL to solve the operation optimization problem of a smart energy hub. Y. Li, Tang, Lv, Guo, and Xu (2018) focused on the dynamic dispatch optimization of the combined cooling, heat and power (CCHP) energy system with uncertain renewable energy and demands. The optimization problem was described as a MDP, and Q-learning was adopted to obtain the optimal or suboptimal policy. H. Liu, Li, Ge, Zhang, and Chen (2019) proposed an optimal scheduling method based on multi-agent game and RL for different investors and operation subjects in the IES.

The energy storage systems (ESSs) in sustainable energy and electric systems have flexibility, which helps to achieve a balance between energy production and energy demand in sustainable energy and electric systems. Specifically, when the energy production of the system exceeds the energy demand or the energy price is low, the ESS can store the surplus energy for subsequent use; when the energy production cannot meet the energy demand or the energy price is high, the ESS can release energy, so as to improve the economy and reliability of the system operation. RL based operation strategies have recently been applied for optimal operation of ESSs in sustainable energy and electric systems. In (Bui, Hussain, & Kim, 2019), DDQN was used to optimize the operation of a community battery ESS in a microgrid. The proposed operation strategy can deal with uncertainties in the system in both grid-connected and islanded modes and maximize the profit in normal mode and minimize load shedding in islanded mode. For integrating multiple types of storage, there are challenges in managing each storage due to its different characteristics. In (Z. Zhang, Qiu, Zhang, Xu, & He, 2019), a hybrid energy system consisting of hydrogen (long-term ESS) and battery (short-term ESS) was constructed. Coordinated control of the hybrid ESSs can be formulated as a sequential decision-making problem and solved effectively with DDQN. Besides, a microgrid energy storage system including lead-acid battery and vanadium redox battery (VRB) was constructed in (X. Qiu, Nguyen, & Crow, 2016), and RL was used to optimize the coordination of different ESSs.

5.1.2. Operation optimization of the user side

Demand side management (DSM) can make users actively participate in the optimal operation of sustainable energy and electric systems, and is able to reduce energy consumption/cost while ensuring energy quality, which is an important means to realize the optimization of user side operation. The implementation of DSM

greatly depends on incorporating user feedback and energy consumption patterns in its control loop. One of the important characteristics of RL is that it is easy to acquire the feedback information of users and learn self-improvement from it. It is a learning method with potential to achieve the goal of user side operation optimization of sustainable energy and electric systems. At present, energy equipment/systems with great potential for DSM applications mainly include domestic hot water (DHW), heating ventilation and air conditioning (HVAC), smart home appliances and electric vehicles (EVs). These energy equipment/systems will not only impact the sustainable energy and electric systems, but also are closely related to human comfort and economic cost. Therefore, it is of great significance to manage them so as to achieve the operation optimization of the user side.

Electric water heaters can store energy in their water buffer without impacting the comfort of the user. This feature makes them a prime candidate for residential demand side management. However, the stochastic and non-linear dynamics of electric water heaters makes it challenging to harness their flexibility. Driven by this challenge, the sequential decision-making problem of electric water heaters can be formulated as a MDP and RL can be used to solve the optimal control strategy (Al-Jabery, Wunsch, Xiong, & Shi, 2014; Al-Jabery, et al., 2017; Ruelens, et al., 2018). In order to mitigate the curse of dimensionality, Ruelens, et al. (2018) applied an automatic encoder network to find a compact feature representation of the sensor measurements, and realized the optimal management of the electric water heater.

HVAC systems can promote the economic operation of sustainable energy and electric systems by pre-heating or pre-cooling the indoor spaces to achieve some degree of load shifting. A lot of research has been carried out on optimizing HVAC operation based on RL (Yujiao Chen, Norford, Samuelson, & Malkawi, 2018; Sun, Luh, Jia, & Yan, 2015; Y. Wang, Velswamy, & Huang, 2017). Y. Wang, et al. (2017) used model-free actor-critic RL to optimize the operation of HVAC in the office space. In order to learn the temporal correlations in temperature state observations, long short-term memory (LSTM) recurrent neural networks (RNNs) were introduced to represent policy and value, which improved energy efficiency without sacrificing the thermal comfort. In addition, Wei, Zhang, Qiao, and Qu (2015) proposed a novel dual iterative Q-learning algorithm to optimize the energy management of the battery installed in the residential user side of residential energy systems, so as to minimize the total expense of the power from the grid and extend the life of the battery.

RL was suggested for demand side management for residential customers in IESs in (Sheikhi, Rayati, & Ranjbar, 2016). The proposed fully automated energy management system (EMS) estimates the residential customers' satisfaction function, energy prices, and efficiencies of appliances based on customers' historical actions, so as to motivate customers for participating in DSM programs and reducing the peak load in both electricity and natural gas networks. Additionally, an optimal operation problem of thermostatically controlled loads (e.g. hot water storage tanks) connected to a district heating network was considered as a sequential decision-making problem under uncertainty in (Claessens, Vanhoudt, Desmedt, & Ruelens, 2018). The decision-making problem was formalized as a MDP and a combination of batch RL (BRL) and multi-agent system was used to minimize the peak demand and reduce the energy cost.

In terms of user side operation optimization of smart buildings, Mocanu, et al. (2019) considerd three types of electrical devices in residential buildings, including time-scaling load (air conditioning), time-shifting load (dishwasher) and time scaling and shifting load (EV). The building environment was modeled as a MDP, and value-based method DQN and policy-based method DPG were respectively used to learn the optimization strategy of demand side operation, so as to optimize building energy consumption or energy cost in an on-line manner. The proposed online operation optimization strategy can cope with the high uncertainty of the energy consumption patterns of users and provide real-time feedback to users to encourage more efficient use of energy. Considering the uncertainty of future electricity price, an ANN can be used to forecast the future electricity price. On this basis, considering the characteristics and priorities of different home appliances, the appliances can be separated into three main types: non-shiftable, shiftable and controllable loads, and multi-agent RL can be adopted to make optimal decisions for multiple home appliances in a decentralized manner, so as to minimize user energy bills and dissatisfaction costs (Lu, Hong, & Yu, 2019).

5.1.3. Operation optimization of the whole system

In order to realize the cooperative operation optimization of the energy supply side and the user side in sustainable energy and electric systems, collaborative management of resources in the energy supply side and responsive resources in the user side can be carried out so as to realize the optimization of their own profit. However, for every supplier, a lack of information about users and other suppliers creates challenges to optimal decision making in order to maximize its return. Similarly, users face difficulty in scheduling their energy consumption without any information about suppliers and energy prices. Additionally, there are several uncertainties involved in energy systems due to fluctuation of renewable energy generation and variability of users' consumption. In order to cope with these challenges, the energy suppliers and users can be modeled as autonomous agents, and RL can be used to realize the overall cooperative operation optimization of the system. When the agent changes its behavior, it can update the model locally based on the new information and learn its new best response without the need for excessive communication to a central controller or other agents, so that the energy suppliers, distributed storage, and users can develop optimal strategies for energy management and load scheduling without prior information about each other and the energy system (Foruzan, Soh, & Asgarpoor, 2018). The energy and load management method consists of five types of agents: wind resources, PV, diesel generators, battery storages, and customers. Interactions of agents and the environment during the learning process lead the system to asymptotically converge to the Nash equilibrium.

5.2. Applications of RL in sustainable energy and electric systems control

With the development and evolution of sustainable energy and electric systems, the system structures and components are changing constantly. There are many unstable factors in the system, and the actual operating conditions are changing all the time. The uncertainty and complexity of system operation are gradually increasing, which requires more robust and adaptive control technologies. Traditional control methods are mostly based on physical characteristics of the system. However, physical modeling is conducted according to different topological structures, operation modes and fault types. Therefore, it is not adaptable to changes of sustainable energy and electric systems structures and new power electronic devices, and it is difficult to meet the development needs of the system. Therefore, researchers try to replace process simulation with data-driven methods, and directly formulate strategies to control the system with the help of autonomous decision-making ability of RL. Power system control is the top priority of sustainable energy and electric systems control, so we mainly focus on the applications of RL in power system control. The power system faces different control problems in different operation states. According to the operation states of the system, the

Table 2
Summary of the applications of RL in power system control.

Classification	Learning algorithm	Reference(s)
Emergency control	RL	Ernst et al. (2004), Glavic (2005), Yu and Zhen (2009), Guo et al. (2006), Rashidi and Rashidi (2003), Mohagheghi et al. (2006), Hadidi and Jeyasurya (2013), Yousefian and Kamalasadan (2015), Zarrabian et al. (2016)
	Combination of RL and classical control method	Glavic et al. (2005), Ernst et al. (2009), D. Wang et al. (2014), B. Li et al. (1999)
	DRL	Wei Liu et al. (2018), Cao et al. (2019)
Normal control	RL	Yu et al. (2009), Yu et al. (2012), Saikia et al. (2011), Xi et al. (2018), Tang et al. (2015), V. P. Singh et al. (2017), Diao et al. (2015), Tousi et al. (2011), Tan et al. (2016)
	DRL	Yin and Yu (2018), Yin et al. (2018)
	Transfer RL	X. Zhang et al. (2017)
MPPT control	RL	C. Wei et al. (2015), Hsu et al. (2015), Kofinas et al. (2017)
	Combination of ANN and RL	C. Wei et al. (2016)

control problems of the power system are divided into security and stability control under emergency state and frequency/voltage control under normal state. Table 2 summarizes the applications of RL in power system control.

5.2.1. Security and stability control of the system under emergency state

The security and stability control of power systems mainly considers that the normal operation of the system can be guaranteed by taking control measures after disturbance. In the case of transient angle instability or oscillatory angle instability, Ernst, Glavic, and Wehenkel (2004) designed a power system stability control framework based on RL which included two modes. The online mode interacted with the real power system and the offline mode interacted with a simulation model of the real power system. The online mode can cope with the difficulty of system modeling or some phenomena cannot be reproduced in a simulation environment. Based on the offline control mode, Glavic (2005) designed a thyristor-switched resistive brake controller with multiple switches. In addition, based on RL algorithms, scholars have designed power system security and stability devices such as power system stabilizer (PSS) (Yu & Zhen, 2009), direct current supplementary damping controller (Guo, Zhang, & Hu, 2006), dynamic quadrature booster (B. Li, Wu, Wang, & Zhou, 1999), static var compensator (SVC) (Rashidi & Rashidi, 2003) and static compensator (Mohagheghi, Venayagamoorthy, & Harley, 2006), which have good adaptability to the stability and dynamics of the system and are suitable for power systems with many uncertainties and severe disturbances.

A more in depth consideration of ways to combine RL with classical control theory methods would be especially valuable in cases of power systems stability control applications. Glavic, Ernst, and Wehenkel (2005) combined control Lyapunov functions (CLF) and online RL control mode built and realized RL optimal control in stable regions. The combination of a stability-oriented and a performance-oriented control technique provides a promising way to implement advanced control schemes in power systems. Moreover, some scholars have proposed to integrate RL methods with existing control technologies such as model predictive control (Ernst, Glavic, Capitanescu, & Wehenkel, 2009; D. Wang, Glavic, & Wehenkel, 2014), fuzzy logic control (B. Li, et al., 1999), so as to achieve a better control effect.

RL has also been involved in wide-area stabilizing control of power systems. In order to stabilize the system after severe disturbances and mitigate the oscillations afterward, a real-time close-loop wide-area decentralized control structure was designed based on a RL method called Least Worst Action Q-learning (Hadidi & Jeyasurya, 2013). In this method, instead of relying on a mathematical model of the system, data from real-time measurements were used. It enlarged the stability domain of the power system and also enhanced the damping of the system oscillations. Yousefian and Kamalasadan (2015) presentd a new method for designing and implementing coordinated wide area controller architecture based on reinforcement and TD learning which allowed the system to learn from interaction and predict future states. The main advantage of this design is its ability to learn from the past using eligibility traces and predict the optimal trajectory through TD learning in the format of receding horizon control.

In terms of congestion control, Zarrabian, Belkacemi, and Babalola (2016) proposed a method based on the RL for preventing cascading failure and blackout in smart grids by acting on the output power of the generators in real-time. It utilizes Q-learning to train the system for the optimal action selection strategy during the state-action learning process by updating the action values based on the obtained rewards. The trained system can relieve congestion of transmission lines in real-time by adjusting the output power of the generators (actions) to prevent consecutive line outages and blackout after N-1 and N-1-1 contingency conditions.

Note that the ability of RL in information perception is weak. In order to improve the correctness of decision-making and control efficiency, the advantages of DL in feature extraction can be taken to extract the operation characteristics in the early stage of analyzing the power grid environmental information and then high-value density information can be provided as input data for RL. Deep CNN and Q-learning were used for power system generating unit tripping strategy in (Wei Liu, Zhang, Wang, Hou, & Liu, 2018), which realized the direct mapping from power system operation information to generating unit tripping control strategy. Aiming at the decision-making optimization problem of transient voltage stability in EI scenario, a DRL algorithm was applied to optimize the reactive power compensation decision of EI (Cao, Zhang, Xiao, & Hua, 2019). The proposed method makes full use of real-time state information and network topology information obtained by massive data acquisition devices and fulfils the expectation of distributed reactive compensation and minimization of total reactive compensation.

5.2.2. Frequency/voltage control of the system under normal state

RL, as a type of self-learning intelligent control algorithm, can be used to solve complex nonlinear frequency/voltage control problems in power systems. The power system frequency is maintained to its nominal value by nullifying the power mismatch between active power generation and load demand using automatic generation control (AGC) (Pathak, Verma, Bhatti, & Nasiruddin, 2019). The AGC is a dynamic multistage decision-making problem and its control process can be formulated as a MDP. Q-learning method based on MDP can be applied to AGC, which regards the control performance standard (CPS) values as the reward from the interconnected power grids. By regulating a closed-loop CPS control rule to maximize the total reward in the procedure of interac-

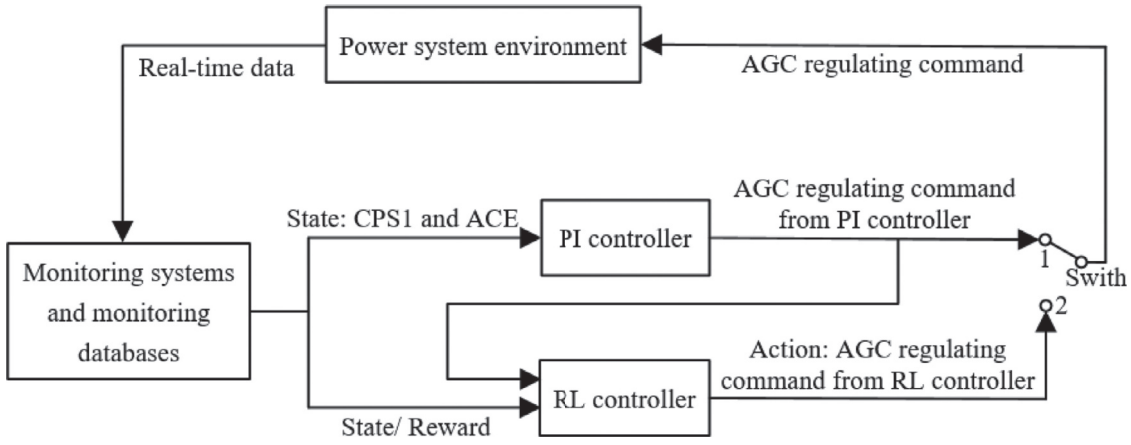

Fig. 6. The imitation pre-learning process of RL controller (adopted and modified from Yu et al., 2019).

tive learning, the robustness and adaptability of the system can be obviously enhanced (Yu, Zhou, & Chan, 2009).

In Yu, et al. (2012), a R(λ) imitation learning (R(λ)IL) method was used to develop an optimal automatic generation controller under CPS. It can learn the control behaviors of the existing automatic generation controller by observing the system changes during the run time. This addresses the defect in the applicability of conventional RL controllers, which an accurate power system model is required for going throught the offline pre-learning process. As shown in Fig. 6, the RL controller observes AGC regulating commands from the existing PI controller to learn and imitate its control behaviors. The system state constists of 1 min moving averages of CPS1 and area control error (ACE), the agent action is the AGC regulating command and the reward function considers both the CPS performance and the relaxed control objective. The pre-learning process will complete once the set termination criterion is satisfied and the toggle switch can be actuated to contactor 2 to deactivate the PI controller and put the RL controller online for live operation.

Furthermore, RL is applied to the automatic generation control of a multi-area hydrothermal hybrid system considering reheat steam turbines in thermal areas and electric governor in hydro areas. The designed controllers are quite robust to wide changes in system loading, inertia constant and size and location of step load perturbation (Saikia, Mishra, Sinha, & Nanda, 2011). Meanwhile, it is noticed that RL is weak in information perception. The prediction mechanism of a DNN can be used as the action selection mechanism of RL to enhance the cognitive ability of the algorithm to the system, so that it can be applied to the robust intelligent power generation control (Yin & Yu, 2018; Yin, Yu, & Zhou, 2018).

Recently, the diversification of distributed generation access to the power grid and increasing randomness of intermittent new energy power generation have made system control more difficult. In order to solve the stochastic disturbance to the power grid caused by the large-scale integration of distributed energy sources, Xi, et al. (2018) proposed a multi-agent RL algorithm named Policy Dynamics based Win or Learn Fast Policy Hill-Climbing (λ) based on the idea of a time tunnel. The optimal strategy can be obtained by adopting the variable learning rate in a variety of complex operating environments, which is difficult for traditional centralized AGC.

Increasing deployment of intermittent power generation in the smart grid will cause large system frequency fluctuation when the load frequency control (LFC) capacity is insufficient to compensate the unbalance of generation and load demand. What is worse, the system inertia will decrease when the smart grid is in islanded mode, which would degrade system damping and cause system instability. At the same time, EVs will be widely used in the near future, and an EV station could be treated as a dispersed battery energy storage station. Therefore, Tang, Yang, Yan, and He (2015) employed the vehicle-to-grid (V2G) technology to compensate for inadequate LFC capacity, thus improving the island grid frequency stability. An on-line RL based method, named goal representation ADP (GrADP), was employed to adaptively control units in an island grid, which held superior instant learning ability and robust control effect. LFC differs from AGC in that AGC includes LFC together with generation dispatch function for control of so called area control error that is a parameterized sum of frequency deviation and active power flows over so-called tie-lines (Glavic, 2019).

Besides, multi-area power systems are impacted by changes in communication topologies (CTs). Their operating mode and power output are often uncontrollable because they are subjected to changes of CTs. As a result, some areas may connect into networks or disconnect from networks at any time. Therefore, the controller design in smart grid environment must have a large degree of intelligence and flexibility to face CT and uncertain environment. Based on RL, V. P. Singh, Kishor, and Samuel (2017) presented an intelligent controller for LFC in smart grid with CT changes using an MAS technique. The controller is composed of an estimator agent and a controller agent. The estimator agent is responsible for estimating the frequency-bias coefficients and the controller agent is used to compensate for the power imbalance between generations against the load demand, which improves the dynamic performance of the system under the LFC scheme.

In power system voltage control, voltage and reactive power control has the inherent characteristics of non-linearity, strong coupling and real-time needs. Its control laws are affected by numerous factors, such as system time variation, changing operating conditions and network parameters. RL methods have strong learning ability and high adaptability, making them very suitable for voltage and reactive power control. Diao, Yang, Chen, and Sun (2015) used Q-learning algorithm to learn continuously under interaction between the action policies and grid states, then got the Q-value function corresponding to each state-action, and finally formed the optimal regional grid reactive power and voltage control strategies. Tousi, Hosseinian, and Menhaj (2011) used a combination of multi-agent system (MAS) technology and RL to provide a desirable voltage profile for the power system. In the proposed schema, individual agents who are assigned to voltage controller devices learn from their experiences to control the system voltage, and also cooperate and communicate with each other to satisfy the whole team goals. Moreover, in order to solve multi-area decentralized collaborative reactive power optimization problem, a hierarchically correlated equilibrium Q-learning (HCEQ) al-

Table 3
The environment state, agent action and reward definition for MPPT control of the WECS.

	Definition
State	(The generator rotor speed, the electrical output power of the wind turbine)
Action	The speed control command: {increment, stay, decrement}
Reward	If the selected action leads to an increment of electrical output power: +1 If the selected action leads to an decrement of electrical output power: −1 If the selected action only leads to a small difference of output power: 0

gorithm for reactive power optimization that considered carbon emission on the grid-side as an optimization objective was proposed by Tan, Han, Zhang, Guo, and Yu (2016). For the reactive power optimization in large-scale power systems, X. Zhang, Yu, Yang, and Cheng (2017) proposed a novel accelerating bio-inspired optimizer (ABO) associated with transfer RL (TRL). A memory matrix was employed to represent the memory of different state-action pairs, which was used for knowledge learning, storage, and transfer among different optimization tasks. The proposed method can rapidly seek the closest solution to the exact global optimum by exploiting the prior knowledge of the source tasks according to their similarities. With the combination of multi-agent system, DL and transfer learning, RL will play an increasingly important role in voltage optimization control.

5.2.3. Clean energy (wind power/PV) maximum power point tracking control

In recent years, wind power has been extensively developed and is expected to be a major alternative source for clean and renewable electricity supply in the future. Currently, the majority of wind power is generated by variable-speed wind energy conversion systems (WECSs). In a variable-speed WECS, the shaft rotating speed can be controlled such that the WECS can track the maximum power points (MPPs) to generate the maximum power for all wind speed conditions. Therefore, an effective MPP tracking (MPPT) control algorithm forms an essential part of the control system of a modern variable-speed WECS. MPPT control methods mainly include tip speed ratio (TSR) control, optimal relationship-based (ORB) control and perturbation and observation (P&O) control. TSR and ORB methods require prior knowledge of wind speed sensors, WECS or cumbersome off-line design process; while in P&O control, WECS will not learn lessons, that is, it has been searching for MPP, or even experienced MPP before, so it responds slowly to changes in wind speed.

In order to overcome the shortcomings of traditional MPPT control methods, which need prior knowledge or cannot be learned from experience, model-free RL is introduced into MPPT control. Generator rotor speed and electrical output power are used as state space and the change of speed control command is used as action space for the controller. A positive reward will be given to the agent if the selected action leads to an increment of the output power, whereas a negative reward will be given if the output power decreases. The controller updates the value of the action according to the received reward and learns the mapping from the state to optimal action online, so as to obtain a general MPPT control method independent of photovoltaic characteristics. Wei, et al. (2015) proposed an intelligent MPPT algorithm for variable-speed WECSs based on the RL method. The definitions of the environmental state, agent action and reward for MPPT control of the WECS are shown in Table 3. The proposed MPPT algorithm consists of two processes: an online learning process and an online application process. In the online learning process, the controller of the WECS behaves similarly to an agent to interact with the environment to learn the MPPs from its own experience using Q-learning. The optimum rotor speed-electrical output power curve is then obtained from these MPPs and used for fast MPPT control of the WECS in the application process. The proposed MPPT method does not require prior knowledge of the WECS and enables the agent to reach the MPP faster at a wind speed that has been experienced before and respond quickly to wind speed variations after learning.

Furthermore, in order to eliminate the need for a large lookup table in the traditional RL whose size is usually selected through offline simulation studies, Wei, Zhang, Qiao, and Qu (2016) combined ANN and Q-learning method for MPPT control of variable-speed WECSs of permanent magnet synchronous generator (PMSG). It not only provides predictions for the knowledge never previously experienced, which shortens the online learning time and improves the learning efficiency, but also saves the computational cost.

As one of the potential new energy sources, solar energy is mainly used for photovoltaic power generation. In order to improve efficiency, a photovoltaic must continuously generate the maximum possible power under different environmental conditions. The MPPT photovoltaic process was defined as a MDP model and an RL algorithm was used to solve the MPPT control problem in (Hsu, Liu, Chen, Hsieh, & Wang, 2015). On this basis, the application scenario of MPPT controller was extended from varying irradiance and stationary temperature to different operating conditions, varying temperature, irradiance and electrical load in (Kofinas, Doltsinis, Dounis, & Vouros, 2017).

5.3. Applications of RL in energy markets

With the increasing social demand for energy, the demand for energy (electricity) as a commodity for free trade is increasing. The research on the energy market of electricity, gas and heat, as well as the impact of energy prices on system operation, has attracted wide attention. Since RL is essentially the pursuit of the maximum cumulative value of long-term reward for an agent, it is very suitable for the decision-making needs of trading individuals in energy markets.

5.3.1. Bidding strategy formulation

In terms of market transactions and bidding strategies, due to the influence of the market operation mode, energy supply and demand and many other factors, the game between the operation of production, transmission, transfer, utilization and storage and the user's energy consumption should be considered, that is, energy transaction equilibrium and bidding strategy in terms of electricity, gas, heat, oil and other energies in a fully competitive market environment. For example, in the bidding process of power suppliers, due to the incomplete information between competitors and the complexity of the trading market, the bidding decision-making process of power suppliers (generators) and their interactions in the market are complex and dynamic problems that are difficult to model explicitly through traditional analytical methods, especially when considering medium and long-term transactions. The RL method is very suitable for analyzing the dynamic behavior of complex systems with uncertainties. It can be used to identify the optimal bidding strategy in the energy market (Kozan, Zlatar, Paravan, & Gubina, 2014). Ma, Jiang, Hou, Bompard, and Wang (2006) proposed a decision-making process model for medium and long-term transactions which could simulate the

strategic bidding of power suppliers based on the Watkins's Q(λ). It does not need an explicit mathematical expression model in the actual optimization process, and can be used to define the optimal bidding strategy for each supplier as well as finding the market equilibrium. Aliabadi, Kaya, and Sahin (2017) studied bidding strategies that consider generation companies' rivals' as well as their own learning behavior and risk aversion levels. To model the learning behavior of generation companies, the researchers used a modified Q-learning algorithm in which each company (agent) experimented with bid price alternatives and learned through experience. Considering the fuzzy nature of human's decision-making processes, a fuzzy system was designed to map each agent's market power into the Q-learning parameters (Rahimiyan & Mashhadi, 2010). The fuzzy Q-learning selects the power supplier's bidding strategy according to the past experiences and the values of the parameters, which show the human risk characteristics, making each agent adapt its response to market changes such as changes in the network topology, generation configuration and rivals' behavior rapidly and profitably.

Methods mentioned above need to discrete the action space, which lead to the information loss since the bidding volume is continuous. Zhao, Wang, Guo, Zhao and Zhang (2016) applied a gradient descent continuous Actor-Critic algorithm in the double-side day-ahead electricity market modeling and simulation. The proposed approach can cope with the issues with continuous state and action sets without causing trouble of curse of dimensionality. Ye, Qiu, Li, and Strbac (2019) proposed a multi-agent DRL based methodology, combining multi-agent intelligence and a DPG-LSTM method, to expedite practical multi-period and multi-spatial equilibrium analysis. In the proposed approach, strategic generation companies (agents) do not rely on any knowledge of the computational algorithm of the market clearing process (environment) and the operating parameters and offering strategies of their competitors, but only on their own operating parameters, the observed market clearing outcomes and the publicly available information on the market condition. The agents are capable of learning their optimal strategies by utilizing experiences acquired from daily repeated interactions with the market clearing process and devising more profitable bidding decisions by exploiting the entire action domain. Considering the physical nonconvex operating characteristics of the electricity producers in the complex bidding markets, an improved DDPG-based methodology was developed to optimizing the strategic bidding decisions in Ye, Qiu, Sun, Papadaskalopoulos, & Strbac (2020).

In terms of renewable energy bidding problem, Li and Shi (2012) proposed to use RL algorithm for the wind generation company bidding optimization so as to maximize its net earnings in the day-ahead electricity markets. Cao, et al. (2020) investigated the possible opportunities for wind power producers if they can take part in the reserve market to schedule some reserve to reduce the risk of being punished due to the inaccuracy of wind power prediction and the uncertainty of electricity price. This problem was formulated as a MDP and A3C was used for the bidding strategy formulation.

5.3.2. Dynamic pricing

In regional energy systems, the service provider (SP) acts as a broker between the utility company and customers by purchasing energy from the utility company and selling it to the customers (CUs). For the SP, although dynamic pricing is an effective tool for managing energy systems, the implementation of dynamic pricing is highly challenging due to the lack of CU-side information and various types of uncertainties in the system. Similarly, the CUs also face challenges in scheduling their energy consumption due to the uncertainty of the retail energy price.

In order to overcome the challenges of implementing dynamic pricing and energy consumption scheduling, Kim, Zhang, Van Der Schaar, and Lee (2016) developed RL algorithms that allow each of the SPs and the CUs to learn its strategy without a priori information about the energy system, which significantly reduced the CUs' cost as well as the system cost. In addition, Lu, Hong, and Zhang (2018) comprehensively considered SP profit and CU costs and formulated the dynamic pricing problem as a discrete finite MDP, and then adopted Q-learning to solve the decision-making problem. The proposed algorithm can promote SP profitability, reduce energy costs for CUs, and balance energy supply and demand in the electricity market, which can be regarded as a win-win strategy for both SP and CUs.

Existing works are mostly concerned with either the bidding problem or the pricing problem. The bidding problem and the pricing problem are inherently coupled, since the energy purchased in the wholesale electricity market and that sold in the retailed electricity market must balance, and the profit earned by the LSE is dependent on the results in both markets. Therefore, it is indeed more desirable to solve the bidding problem and the pricing problem jointly. Xu, Sun, Nikovski, Kitamura, Mori, and Hashimoto (2019) studied the problem of jointly determining the energy bid submitted to the wholesale electricity market and the energy price charged in the retailed electricity market for a load serving entity. The joint bidding and pricing problem is formulated as a MDP formulation with continuous state and action spaces, in which the energy bid and the energy price are two actions that share a common objective. DDPG was applied to solve this MDP for the optimal bidding and pricing policies.

In addiation, Chen and Su (2019) explored the role of emerging energy brokers (middlemen) in a localized event-driven market at the distribution level for facilitating indirect customer-to- customer energy trading. The energy trading process was built as a MDP and a modified Q-Learning algorithm was used to solve this problem. In Lincoln, Galloway, Stephen and Burt (2012), RL algorithms were applied in simulations of competitive electricity trade. In this paper, two policy gradient algorithms using ANN for policy function approximation were compared with a value-function based method in simulations of electricity trade.

5.4. Applications of RL in cyber security and defense

Modern sustainable energy and electric systems are no longer traditional physical systems, but have envolved into cyber and physical deep coupling energy cyber physical systems (D. Liu, Sheng, Wang, Lu, & Sun, 2015; Yang, Huang, Pen, & Zhang, 2017; Yang, Zhang, Li, & Zomaya, 2020). The weak link of the system is no longer confined to the physical subsystem, but the cyber link may become a new weak spot. Therefore, the cyber security and defense of energy cyber physical systems becomes more and more important. The 2010 Iranian nuclear power earthquake network attack and the 2013 Ukrainian blackout are more factual reminders of the importance of network security in the national energy system.

Intrusion detection is the primary defense technology in cyber security. At present, the known intrusion attack methods for energy cyber physical systems mainly include false data injection, virus attack, black hole attack, eavesdropping attack, denial of service (DoS) attack and so on (Mo, et al., 2011). Moreover, they are moving towards the trend of teamwork combined attack.

5.4.1. From the perspective of the defender

Recently, researchers have introduced RL into the process of power intrusion detection. Some established intrusion detection system models based on semi-MDP (SMDP) RL that can detect not only known intrusions, but also unknown intrusions

(S. Li, Wang, Wang, & Niu, 2006), or formulated the online cyberattack/anomaly detection problem as a partial observable MDP (POMDP) problem that can be solved by model-free RL (Kurt, Ogundijo, Li, & Wang, 2019). In order to minimize the cyberattack impacts on power infrastructures, a DRL based recovery strategy was proposed to optimally reclose the transmission lines lost in the cyberattack (Wei, Wan, & He, 2019). The DRL framework endowed the recovery strategy with environmental adaptability and real-time decision-making ability.

5.4.2. From the perspective of the attacker

The above research considers the system security from the perspective of the defender. In addition, some scholars have studied it from the perspective of the attacker, whose aim is to determine the attack strategy that leads to maximized damage to the system, analyze the vulnerability of the system and take necessary precautions accordingly. Yan, He, Zhong, and Tang (2016) used a Q-learning based approach to adaptively identify the more vulnerable attack sequence that could cause critical system failure from sequential topology attacks based on the self-learning ability of RL. Ying Chen, Huang, Liu, Wang, and Sun (2019) modeled the optimal attack strategy as a POMDP and adopted a Q-learning algorithm with nearest sequence memory to enable on-line false data injection attacks on power systems. RL can also be adopted to study the multistage dynamic game between the attacker and the defender. Given a certain objective (e.g., transmission line outages or generation loss), the optimal attack sequences can be identified. Through the learning of "simulated" attackers, defenders can eventually better prepare defense strategies (Ni & Paul, 2019).

Moreover, by seamlessly combining RL with game theory and fuzzy cluster based analytical methods, security situational awareness for the smart grid can be achieved (Jun Wu, Ota, Dong, Li, & Wang, 2018). Based on the proposed mechanism, the extraction of cyber security situation factors, cyber situational assessment and security situational prediction can be realized for the smart grid.

5.5. Applications of RL in electric vehicle management

With the increasingly serious global energy shortage and deepening environmental problems, EVs are favored by countries all over the world due to advantages of high efficiency, energy saving and economy, and have become a research hotspot. However, the access of EVs to the power grid will not only affect the load and power quality of distribution networks, but also affect the reliability and stable operation of the grid. Especially when the penetration of EVs is high, the threat will become more apparent if the charging of EV is not guided. The charging strategy of EV is a dynamic decision-making problem, which can be modeled as a MDP and then solved by RL.

Chiş, Lundén, and Koivunen (2017) proposed a batch RL based charging strategy that aimed at reducing the long-term cost of charging the battery of an individual plug-in EV (PEV). The method makes use of actual electricity prices for the current day and predicted electricity prices for the following day which captures the day-to-day differences of electricity charging costs and minimizes the charging cost of PEV. Arif, et al. (2016) studied different dispatching strategies of EVs under different price structures and formulated the scheduling problem of PEV charging as a multistage decision making problem under uncertainty. Finally, Q-learning algorithm was used to find the optimum schedule. Vandael, Claessens, Ernst, Holvoet, and Deconinck (2015) addressed the problem of defining a day-ahead consumption plan for charging a fleet of EVs without an accurate mathematical model of charging flexibility. The charging behavior of the EV fleet was learned by using batch mode RL and a cost-effective day-ahead plan could be defined.

Considering the randomness of both the electricity price and the commuting behavior, Wan, Li, He, and Prokhorov (2019) formulated the EV charging/discharging scheduling problem as a MDP with unknown transition probability. A model-free approach based on DRL was proposed to determine the optimal strategy for the real-time scheduling problem. The architecture of the proposed approach contains a representation network and a Q network. The representation network extracts discriminative features from the electricity price. After concatenating these features with battery SOC, the concatenated features are fed into a Q network to approximate the action-value of all feasible schedules under the given input state. The schedule with the largest action-value is selected as the EV charging/discharging schedule. The overall diagram of the approach is illustrated in Fig. 7. The electricity price trends are captured by a LSTM network. Its input is the past 24-hour electricity price and its output is the features containing information about future price trends. The output of the LSTM network is concatenated with the battery SOC. Then, they are fed into the Q network to approximate the optimal action-value function $Q(s, a)$.

Electric taxis (ETs) are closely related to people's lives, with the characteristics of high driving density and uncertain driving routes. In addition, the operation behavior of ETs involves passengers, transportation network, charging station and power grid, etc., and its decision-making is influenced by many factors, which has strong randomness and dynamics. At the same time, the purpose of ET is profit-making and the ET is limited by the fixed charging time and location. Its operation and load characteristics will be more complex than private EVs and electric buses, and the resulting random load will have a greater impact on the uncertainty of the power grid. Traditional charging load modeling of EVs mostly adopts MC method or mathematical optimization method, which needs to greatly simplify the relevant models and is difficult to consider the autonomy and self-adaptability of each agent. It is difficult to accurately simulate the actual operation system in temporal and spatial scales, especially for plug-in ETs (PETs) with strong randomness and complex operation behavior. In order to solve this problem, Jiang, Jing, Cui, Ji, and Wu (2018) proposed a real-time simulation system for PET operation based on multi-agent technology and RL. In this system, a variety of agent models have been built, such as PET agent, map agent, charging station agent and power grid agent. Multi-step Q(λ) is used to dynamically simulate the decision-making process of PET. The system can find the optimal strategy by adjusting its behavior through constantly interactive learning between agents and the environment.

In addition, in terms of the design of EMS of EV, with the help of self-learning ability of RL, Qi, Wu, Boriboonsomsin, Barth, and Gonder (2016) and Xiong, Cao, and Yu (2018) proposed real-time RL-based EMSs for plug-in hybrid EV, which solved the contradiction between the real-time performance of EV and optimal energy saving. In (T. Liu, Zou, Liu, & Sun, 2015) and (Zou, Liu, Liu, & Sun, 2016), Q-learning was applied to the energy management for a hybrid tracked vehicle, which was more economical than stochastic DP. Moreover, in order to solve the problem of curse of dimensionality resulted by too many discrete inputs of states in RL, DRL was introduced in EV energy management in (Jingda Wu, He, Peng, Li, & Li, 2018) and (Qi, Luo, Wu, Boriboonsomsin, & Barth, 2019). Compared with traditional RL-based strategies, more state variables are considered in DRL which means the agent can obtain and understand environmental information more accurately. The EV energy management strategy obtained is also more accurate.

5.6. Other applications

In smart grid and EI, flexible and efficient communication mechanisms are very important. Software-defined networking (SDN) has been used for real-time monitoring and communicat-

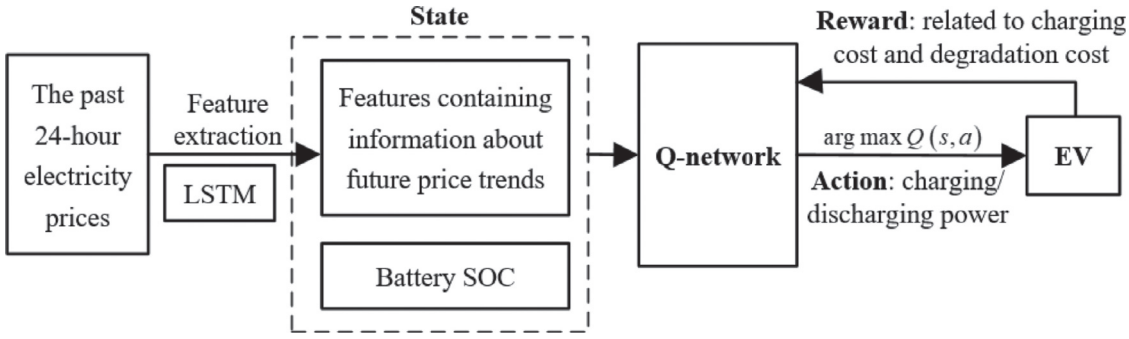

Fig. 7. The Overall diagram of the EV scheduling approach (adopted and modified from Wan, Li, He, & Prokhorov, 2019).

ing in EI. However, how to manage multiple controllers automatically and intelligently in software-defined EI (SDEI) so as to keep high accuracy in the real-time monitoring in complex EI environment is challenging. In distributed SDEI, RL has also been tentatively attempted. A controller mind framework to implement automatic management among multiple controllers based on RL in distributed SDEI was proposed in (C. Qiu, et al., 2019). The problem was described as a MDP, and Q-learning was adopted to identify the optimal scheme to allocate packet-in messages from the data plane to the control plane with the minimum waiting time of QoS flows and the acceptable packet loss rate of best-effort flows.

Based on the characteristics of interactive learning of RL, the application of autonomous learning ability of RL is extended to smart cities and Internet of Things (IoT) ecosystems. Providing a large amount of training data is not always feasible in smart city scenarios. Therefore, we need to consider alternative ways that incorporate unlabeled data as well. Data can be partially labeled by users' feedback for training purposes through RL, which provides a feasible solution for supporting smart city services. Mohammadi, Al-Fuqaha, Guizani, and Oh (2018) proposed a semi-supervised DRL framework as a learning mechanism in support of smart IoT services. The proposed model used labeled data along with unlabeled data to improve the performance and accuracy of the learning agent. Besides, He, Yu, Zhao, Leung, and Yin (2017) jointly considered networking, caching, and computing to enhance the performance of applications for smart cities and formulated the resource allocation strategy as a joint optimization problem. DQN was used to enable dynamic orchestration of networking, caching, and computing resources in applications for smart cities.

Furthermore, the integration of IoT and autonomous control systems results in a new concept: autonomous IoT (AIoT). Lei, Tan, Liu, and Zheng (2019) discussed the applications and challenges of DRL in AIoT based on the structure of the IoT (including perception layer, network layer and application layer). Among them, the applications of DRL in the perception layer include applications in autonomous robots, smart vehicles and smart grid, the applications in the network layer include applications in IoT communications networks, and the applications in the application layer include applications in IoT edge/fog/cloud computing systems.

6. Challenges and prospects

With the rise of artificial intelligence to a strategic issue of many countries around the world, the integration of new generation AI technologies and sustainable energy and electric systems will become closer. However, the research works about applications of RL in sustainable energy and electric systems are largely still in laboratory research stage, which have not been practically implemented in sustainable energy and electric systems. There is still a long way to go from the "usable" stage to the "well used" stage. The following are some challenges and prospects:

1. A key point of using RL is to design an appropriate reward function (Littman, 2015). At present, the reward functions are mainly defined by researchers manually. An unreasonable reward will affect the final optimal policy to a great extent. A so-called inverse RL method addresses the challenge of creating appropriate reward functions given the availability of behavioural traces from an expert executing the target behaviour (Littman, 2015). Foe example, some scholars have tried to improve RL with the help of human teaching, so that the agent can better learn the expected actions (Thomaz & Breazeal, 2008). Therefore, the research on reward functions will be a hot spot in the future development of RL.
2. Systems integrating electric energy, thermal energy, gas and other energy into sustainable energy and electric systems are in the initial stage of development, and the complexity and uncertainty in the systems is more serious than conventional single energy systems, so there is a high demand for RL in this field. At present, research of RL in IESs is still rare, and it can be predicted that the application of RL in IESs will be a fast developing research direction in the future. Similarly, with the increasingly coupling of cyber systems and physical systems in energy systems, the application of RL in energy cyber physical systems (including energy IoT) will also receive more and more attention (Glavic, 2019).
3. In terms of energy forecasting (including renewable energy power forecasting and energy load forecasting), most of the widely used machine learning models are trained by optimizing the global performance, without considering the local behavior. Feng, Sun, and Zhang (2020) introduced RL into power load forecasting, and adopted RL to select the most appropriate method from a number of available forecasting methods, which provided a good idea for the application of RL in the field of energy forecasting (Glavic, 2019). In addition, the method can also be extended to energy price forecasting in the energy market.
4. Transfer learning (Pan & Yang, 2010) is a popular machine learning technology recently. It can learn the general structure of data in a data-driven way and is regared as an important research direction of AI in the future. However, its applications in the energy field are less reported at present. It is a new research topic worth exploring to introduce transfer learning into RL, so as to leverage the advantages of transfer learning to better serve sustainable energy and electric systems.
5. The high-level application of RL in sustainable energy and electric systems is the deep integration of intelligent sensor, big data analysis, machine learning, natural language processing and other technologies, so as to evolve into intelligent machines. Although there are power robots for the application scenarios such as power inspection and equipment live mainte-

nance at present, they do not have autonomous intelligent behavior and can only complete a single mode operation according to the set procedure. Therefore, breaking through the key technologies such as autonomous learning and autonomous behavior, and developing intelligent power robots with humanoid behavior ability will be a future development direction.

6. By embedding prior knowledge of specific fields into RL, human intelligence and artificial intelligence are fused to form a higher and stronger level of intelligence than each of them individually, which will also has a broad application prospect in the field of sustainable energy and electric systems operation and control.

Conflict of Interest

No potential conflict of interest exists for this article.

Acknowledgements

This effort was sponsored by National natural science foundation of China (61971305, 61571324), National key research and development plan (2017YFE0132100), and Natural Science Foundation of Tianjin, China (19JCQNJC06000) .

References

Al-Jabery, K., Wunsch, D. C., Xiong, J., & Shi, Y. (2014). A novel grid load management technique using electric water heaters and Q-learning. *International Conference on Smart Grid Communications*, 776–781.

Al-Jabery, K., Xu, Z., Yu, W., Wunsch, D. C., Xiong, J., & Shi, Y. (2017). Demand-side management of domestic electric water heaters using approximate dynamic programming. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 36*(5), 775–788.

Aliabadi, D. E., Kaya, M., & Sahin, G. (2017). Competition, risk and learning in electricity markets: An agent-based simulation study. *Applied energy, 195*, 1000–1011.

Arif, A., Babar, M., Ahamed, T. I., Al-Ammar, E., Nguyen, P., Kamphuis, I. R., & Malik, N. (2016). Online scheduling of plug-in vehicles in dynamic pricing schemes. *Sustainable Energy, Grids and Networks, 7*, 25–36.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine, 34*(6), 26–38.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics, 13*(5), 834–846.

Bu, L., Babu, R., & De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38*(2), 156–172.

Bui, V.-H., Hussain, A., & Kim, H.-M. (2020). Double deep Q-learning-based distributed operation of battery energy storage system considering uncertainties. *IEEE Transactions on Smart Grid, 11*(1), 457–469.

Buşoniu, L., de Bruin, T., Tolić, D., Kober, J., & Palunko, I. (2018). Reinforcement learning for control: Performance, stability, and deep approximators. *Annual Reviews in Control, 46*, 8–28.

Cao, D., Hu, W., Xu, X., Dragičević, T., Huang, Q., Liu, Z., Chen, Z., & Blaabjerg, F. (2020). Bidding strategy for trading wind energy and purchasing reserve of wind power producer–A DRL based approach. *International Journal of Electrical Power & Energy Systems, 117*, 105648.

Cao, J., Zhang, W., Xiao, Z., & Hua, H. (2019). Reactive power optimization for transient voltage stability in energy internet via deep reinforcement learning approach. *Energies, 12*(8), 1556.

Chen, T., & Su, W. (2019). Indirect customer-to-customer energy trading with reinforcement learning. *IEEE Transactions on Smart Grid, 10*(4), 4338–4348.

Chen, Y., Huang, S., Liu, F., Wang, Z., & Sun, X. (2019). Evaluation of reinforcement learning-based false data injection attack to automatic voltage control. *IEEE Transactions on Smart Grid, 10*(2), 2158–2169.

Chen, Y., Norford, L. K., Samuelson, H. W., & Malkawi, A. (2018). Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. *Energy and Buildings, 169*, 195–205.

Chiş, A., Lundén, J., & Koivunen, V. (2017). Reinforcement learning-based plug-in electric vehicle charging with forecasted price. *IEEE Transactions on Vehicular Technology, 66*(5), 3674–3684.

Claessens, B. J., Vanhoudt, D., Desmedt, J., & Ruelens, F. (2018). Model-free control of thermostatically controlled loads connected to a district heating network. *Energy and Buildings, 159*, 1–10.

Degris, T., Pilarski, P. M., & Sutton, R. S. (2012). Model-free reinforcement learning with continuous action in practice. *American Control Conference*, 2177–2182.

Diao, H., Yang, M., Chen, F., & Sun, G. (2015). Reactive power and voltage optimization control approach of the regional power grid based on reinforcement learning theory. *Transactions of China Electrotechnical Society, 30*(12), 408–414.

Du, Y., & Li, F. (2020). Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning. *IEEE Transactions on Smart Grid, 11*(2), 1066–1076.

Ernst, D., Glavic, M., Capitanescu, F., & Wehenkel, L. (2009). Reinforcement learning versus model predictive control: A comparison on a power system problem. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39*(2), 517–529.

Ernst, D., Glavic, M., & Wehenkel, L. (2004). Power systems stability control: reinforcement learning framework. *IEEE Transactions on Power Systems, 19*(1), 427–435.

Feng, C., Sun, M., & Zhang, J. (2020). Reinforced deterministic and probabilistic load forecasting via Q-learning dynamic model selection. *IEEE Transactions on Smart Grid, 11*(2), 1377–1386.

Foruzan, E., Soh, L.-K., & Asgarpoor, S. (2018). Reinforcement learning approach for optimal distributed energy management in a microgrid. *IEEE Transactions on Power Systems, 33*(5), 5749–5758.

Glavic, M. (2005). Design of a resistive brake controller for power system stability enhancement using reinforcement learning. *IEEE Transactions on Control Systems Technology, 13*(5), 743–751.

Glavic, M. (2019). (Deep) Reinforcement learning for electric power system control and related problems: A short review and perspectives. *Annual Reviews in Control, 48*, 22–35.

Glavic, M., Ernst, D., & Wehenkel, L. (2005). Combining a stability and a performance-oriented control in power systems. *IEEE Transactions on Power Systems, 20*(1), 525–526.

Guo, L., Zhang, Y., & Hu, J. (2006). An adaptive HVDC supplementary damping controller based on reinforcement learning. *IET International Conference on Advances in Power System Control, Operation and Management*, 149–153.

Hadidi, R., & Jeyasurya, B. (2013). Reinforcement learning based real-time wide-area stabilizing control agents to enhance power system stability. *IEEE Transactions on Smart Grid, 4*(1), 489–497.

He, Y., Yu, F. R., Zhao, N., Leung, V. C., & Yin, H. (2017). Software-defined networks with mobile edge computing and caching for smart cities: A big data deep reinforcement learning approach. *IEEE Communications Magazine, 55*(12), 31–37.

Hinton, G. E. (2012). A practical guide to training restricted Boltzmann machines. In *Neural Networks: Tricks of the Trade* (pp. 599–619). Springer.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18*(7), 1527–1554.

Hsu, R. C., Liu, C.-T., Chen, W.-Y., Hsieh, H.-I., & Wang, H.-L. (2015). A reinforcement learning-based maximum power point tracking method for photovoltaic array. *International Journal of Photoenergy* 2015.

Hu, B., Lu, Z., Li, H., & Chen, Q. (2014). Convolutional neural network architectures for matching natural language sentences. *Advances in Neural Information Processing Systems*, 2042–2050.

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.

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.

Jiang, C., Jing, Z., Cui, X., Ji, T., & Wu, Q. (2018). Multiple agents and reinforcement learning for modelling charging loads of electric taxis. *Applied Energy, 222*, 158–168.

Jiao, R., Huang, X., Ma, X., Han, L., & Tian, W. (2018). A model combining stacked auto encoder and back propagation algorithm for short-term wind power forecasting. *IEEE Access, 6*, 17851–17858.

Jin, X., Mu, Y., Jia, H., Wu, J., Xu, X., & Yu, X. (2016). Optimal day-ahead scheduling of integrated urban energy systems. *Applied Energy, 180*, 1–13.

Khan, S. G., Herrmann, G., Lewis, F. L., Pipe, T., & Melhuish, C. (2012). Reinforcement learning and optimal adaptive control: An overview and implementation examples. *Annual Reviews in Control, 36*(1), 42–59.

Kim, B.-G., Zhang, Y., Van Der Schaar, M., & Lee, J.-W. (2016). Dynamic pricing and energy consumption scheduling with reinforcement learning. *IEEE Transactions on Smart Grid, 7*(5), 2187–2198.

Kofinas, P., Doltsinis, S., Dounis, A., & Vouros, G. (2017). A reinforcement learning approach for MPPT control method of photovoltaic sources. *Renewable Energy, 108*, 461–473.

Kofinas, P., Dounis, A., & Vouros, G. (2018). Fuzzy Q-Learning for multi-agent decentralized energy management in microgrids. *Applied Energy, 219*, 53–67.

Kozan, B., Zlatar, I., Paravan, D., & Gubina, A. (2014). The advanced bidding strategy for power generators based on reinforcement learning. *Energy Sources, Part B: Economics, Planning, and Policy, 9*(1), 79–86.

Kurt, M. N., Ogundijo, O., Li, C., & Wang, X. (2019). Online cyber-attack detection in smart grid: a reinforcement learning approach. *IEEE Transactions on Smart Grid, 10*(5), 5174–5185.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Lei, L., Tan, Y., Liu, S., & Zheng, K. (2019). Deep reinforcement learning for autonomous internet of things: model, applications and challenges. *arXiv preprint arXiv:1907.09059*.

Li, B., Wu, Q., Wang, P., & Zhou, X. (1999). Learning-coordinated fuzzy logic control of dynamic quadrature boosters in multi-machine power systems. *IEE Proceedings-Generation, Transmission and Distribution, 146*(6), 577–585.

Li, G., & Shi, J. (2012). Agent-based modeling for trading wind power with uncer-

tainty in the day-ahead wholesale electricity markets of single-sided auctions. *Applied Energy, 99*, 13–22.
Li, S., Wang, X., Wang, Q., & Niu, S. (2006). Research on intrusion detection based on SMDP reinforcement learning in electric power information network. *Electric Power Automation Equipment, 26*(12), 75–78.
Li, Y., Tang, H., Lv, K., Guo, X., & Xu, D. (2018). Modeling and learning-based optimization of the energy dispatch for a combined cooling, heat and power microgrid system with uncertain sources and loads. *Control Theory & Applications, 35*(1), 56–64.
Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.
Lincoln, R., Galloway, S., Stephen, B., & Burt, G. (2012). Comparing policy gradient and value function based reinforcement learning methods in simulated electrical power trade. *IEEE Transactions on Power Systems, 27*(1), 373–380.
Littman, M. L. (2015). Reinforcement learning improves behaviour from evaluative feedback. *Nature, 521*(7553), 445–451.
Liu, D., Sheng, W., Wang, Y., Lu, Y., & Sun, C. (2015). Key technologies and trends of cyber physical system for power grid. *Proceedings of the CSEE, 35*(14), 3522–3531.
Liu, H., Li, J., Ge, S., Zhang, P., & Chen, X. (2019). Coordinated scheduling of grid-connected integrated energy microgrid based on multi-agent game and reinforcement learning. *Automation of Electric Power Systems, 43*(1), 40–48.
Liu, J., Gao, F., & Luo, X. (2019). Survey of deep reinforcement learning based on value function and policy gradient. *Chinese Journal of Computers, 42*(6), 1406–1438.
Liu, T., Zou, Y., Liu, D., & Sun, F. (2015). Reinforcement learning of adaptive energy management with transition probability for a hybrid electric tracked vehicle. *IEEE Transactions on Industrial Electronics, 62*(12), 7837–7846.
Liu, W., Zhang, D., Wang, X., Hou, J., & Liu, L. (2018). A decision making strategy for generating unit tripping under emergency circumstances based on deep reinforcement learning. *Proceedings of the CSEE, 38*(1), 109–119.
Liu, W., Zhuang, P., Liang, H., Peng, J., & Huang, Z. (2018). Distributed economic dispatch in microgrids based on cooperative reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems, 29*(6), 2192–2203.
Lu, R., Hong, S. H., & Yu, M. (2019). Demand response for home energy management using reinforcement learning and artificial neural network. *IEEE Transactions on Smart Grid, 10*(6), 6629–6639.
Lu, R., Hong, S. H., & Zhang, X. (2018). A dynamic pricing demand response algorithm for smart grid: reinforcement learning approach. *Applied Energy, 220*, 220–230.
Ma, Y., Jiang, C., Hou, Z., Bompard, E., & Wang, C. (2006). Strategic bidding of the electricity producers based on the reinforcement learning. *Proceedings of the CSEE, 26*(17), 12–17.
Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. *International Conference on Machine Learning*, 1928–1937.
Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.
Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., & Ostrovski, G. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533.
Mo, Y., Kim, T. H.-J., Brancik, K., Dickinson, D., Lee, H., Perrig, A., & Sinopoli, B. (2011). Cyber-physical security of a smart grid infrastructure. *Proceedings of the IEEE, 100*(1), 195–209.
Mocanu, E., Mocanu, D. C., Nguyen, P. H., Liotta, A., Webber, M. E., Gibescu, M., & Slootweg, J. G. (2019). On-line building energy optimization using deep reinforcement learning. *IEEE Transactions on Smart Grid, 10*(4), 3698–3708.
Mohagheghi, S., Venayagamoorthy, G. K., & Harley, R. G. (2006). Adaptive critic design based neuro-fuzzy controller for a static compensator in a multimachine power system. *IEEE Transactions on Power Systems, 21*(4), 1745–1755.
Mohammadi, M., Al-Fuqaha, A., Guizani, M., & Oh, J.-S. (2018). Semisupervised deep reinforcement learning in support of IoT and smart city services. *IEEE Internet of Things Journal, 5*(2), 624–635.
Ni, Z., & Paul, S. (2019). A multistage game in smart grid security: A reinforcement learning solution. *IEEE Transactions on Neural Networks and Learning Systems, 30*(9), 2684–2695.
Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359.
Pathak, N., Verma, A., Bhatti, T. S., & Nasiruddin, I. (2019). Modeling of HVDC tie links and their utilization in AGC/ LFC operations of multiarea power systems. *IEEE Transactions on Industrial Electronics, 66*(3), 2185–2197.
Qi, X., Luo, Y., Wu, G., Boriboonsomsin, K., & Barth, M. (2019). Deep reinforcement learning enabled self-learning control for energy efficient driving. *Transportation Research Part C: Emerging Technologies, 99*, 67–81.
Qi, X., Wu, G., Boriboonsomsin, K., Barth, M. J., & Gonder, J. (2016). Data-driven reinforcement learning-based real-time energy management system for plug-in hybrid electric vehicles. *Transportation Research Record, 2572*(1), 1–8.
Qiu, C., Cui, S., Yao, H., Xu, F., Yu, F. R., & Zhao, C. (2019). A novel QoS-enabled load scheduling algorithm based on reinforcement learning in software-defined energy internet. *Future Generation Computer Systems, 92*, 43–51.
Qiu, X., Nguyen, T. A., & Crow, M. L. (2016). Heterogeneous energy storage optimization for microgrids. *IEEE Transactions on Smart Grid, 7*(3), 1453–1461.
Rahimiyan, M., & Mashhadi, H. R. (2010). An adaptive Q-learning algorithm developed for agent-based computational modeling of electricity market. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(5), 547–556.
Rashidi, M., & Rashidi, F. (2003). Damping enhancement in the presence of load parameters uncertainty using reinforcement learning based SVC controller. *International Conference on Systems, Man and Cybernetics*, 3068–3072 pp. 3068-3072.
Rayati, M., Sheikhi, A., & Ranjbar, A. M. (2015). Optimising operational cost of a smart energy hub, the reinforcement learning approach. *International Journal of Parallel, Emergent and Distributed Systems, 30*(4), 325–341.
Recht, B., Re, C., Wright, S., & Niu, F. (2011). Hogwild: A lock-free approach to parallelizing stochastic gradient descent. *Advances in Neural Information Processing Systems*, 693–701.
Rocchetta, R., Bellani, L., Compare, M., Zio, E., & Patelli, E. (2019). A reinforcement learning framework for optimal operation and maintenance of power grids. *Applied Energy, 241*, 291–301.
Ruelens, F., Claessens, B. J., Quaiyum, S., De Schutter, B., Babuška, R., & Belmans, R. (2018). Reinforcement learning applied to an electric water heater: from theory to practice. *IEEE Transactions on Smart Grid, 9*(4), 3792–3800.
Rummery, G. A., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems*. England: University of Cambridge, Department of Engineering Cambridge.
Saikia, L. C., Mishra, S., Sinha, N., & Nanda, J. (2011). Automatic generation control of a multi area hydrothermal system using reinforced learning neural network controller. *International Journal of Electrical Power & Energy Systems, 33*(4), 1101–1108.
Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2016). Prioritized experience replay. *arXiv preprint arXiv:1511.05952*.
Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust region policy optimization. *International Conference on Machine Learning*, 1889–1897 pp. 1889-1897.
Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
Sheikhi, A., Rayati, M., & Ranjbar, A. M. (2016). Demand side management for a residential customer in multi-energy systems. *Sustainable Cities and Society, 22*, 63–77.
Shin, M., Ryu, K., & Jung, M. (2012). Reinforcement learning approach to goal-regulation in a self-evolutionary manufacturing system. *Expert Systems with Applications, 39*(10), 8736–8743.
Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., & Riedmiller, M. (2014). Deterministic policy gradient algorithms. *International Conference on Machine Learning*, 387–395 pp. 387-395.
Singh, S., Jaakkola, T., Littman, M. L., & Szepesvári, C. (2000). Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning, 38*(3), 287–308.
Singh, V. P., Kishor, N., & Samuel, P. (2017). Distributed multi-agent system-based load frequency control for multi-area power system in smart grid. *IEEE Transactions on Industrial Electronics, 64*(6), 5151–5160.
Sun, B., Luh, P. B., Jia, Q., & Yan, B. (2015). Event-based optimization within the Lagrangian relaxation framework for energy savings in HVAC systems. *IEEE Transactions on Automation Science and Engineering, 12*(4), 1396–1406.
Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning, 3*(1), 9–44.
Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.
Sutton, R. S., McAllester, D. A., Singh, S. P., & Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 1057–1063.
Tan, M., Han, C., Zhang, X., Guo, L., & Yu, T. (2016). Hierarchically correlated equilibrium Q-learning for multi-area decentralized collaborative reactive power optimization. *CSEE Journal of Power and Energy Systems, 2*(3), 65–72.
Tang, Y., Yang, J., Yan, J., & He, H. (2015). Intelligent load frequency controller using GrADP for island smart grid with electric vehicles and renewable resources. *Neurocomputing, 170*, 406–416.
Thomaz, A. L., & Breazeal, C. (2008). Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence, 172*(6-7), 716–737.
Tousi, M. R., Hosseinian, S. H., & Menhaj, M. B. (2011). A Multi-agent-based voltage control in power systems using distributed reinforcement learning. *Simulation, 87*(7), 581–599.
van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double Q-learning. *AAAI Conference on Artificial Intelligence*.
van Hasselt, H. (2010). Double Q-learning. *Advances in Neural Information Processing Systems*, 2613–2621.
Vandael, S., Claessens, B., Ernst, D., Holvoet, T., & Deconinck, G. (2015). Reinforcement learning of heuristic EV fleet charging in a day-ahead electricity market. *IEEE Transactions on Smart Grid, 6*(4), 1795–1805.
Venayagamoorthy, G. K., Sharma, R. K., Gautam, P. K., & Ahmadi, A. (2016). Dynamic energy management system for a smart microgrid. *IEEE Transactions on Neural Networks and Learning Systems, 27*(8), 1643–1656.
Wan, Z., Li, H., He, H., & Prokhorov, D. (2019). Model-free real-time EV charging scheduling based on deep reinforcement learning. *IEEE Transactions on Smart Grid, 10*(5), 5246–5257.
Wang, D., Glavic, M., & Wehenkel, L. (2014). Trajectory-based supplementary damping control for power system electromechanical oscillations. *IEEE Transactions on Power Systems, 29*(6), 2835–2845.
Wang, Y., Velswamy, K., & Huang, B. (2017). A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems. *Processes, 5*(3), 46.
Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., & de Fre-

itas, N. (2016). Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*.
Wang, Z., Schaul, T., Hessel, M., Hasselt, H., Lanctot, M., & Freitas, N. (2016). Dueling network architectures for deep reinforcement learning. *International Conference on Machine Learning*, 1995–2003.
Watkins, C. (1989). *Learning from delayed rewards PhD thesis*. King's College, University of Cambridge.
Wei, C., Zhang, Z., Qiao, W., & Qu, L. (2015). Reinforcement-learning-based intelligent maximum power point tracking control for wind energy conversion systems. *IEEE Transactions on Industrial Electronics, 62*(10), 6360–6370.
Wei, C., Zhang, Z., Qiao, W., & Qu, L. (2016). An adaptive network-based reinforcement learning method for MPPT control of PMSG wind energy conversion systems. *IEEE Transactions on Power Electronics, 31*(11), 7837–7848.
Wei, F., Wan, Z., & He, H. (2019). Cyber-attack recovery strategy for smart grid based on deep reinforcement learning. *IEEE Transactions on Smart Grid* Early access, https://ieeexplore.ieee. org/document/ 8915727.
Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning, 8*(3-4), 229–256.
Wu, J., He, H., Peng, J., Li, Y., & Li, Z. (2018). Continuous reinforcement learning of energy management with deep Q network for a power split hybrid electric bus. *Applied Energy, 222*, 799–811.
Wu, J., Ota, K., Dong, M., Li, J., & Wang, H. (2018). Big data analysis-based security situational awareness for smart grid. *IEEE Transactions on Big Data, 4*(3), 408–417.
Xi, L., Chen, J., Huang, Y., Xu, Y., Liu, L., Zhou, Y., & Li, Y. (2018). Smart generation control based on multi-agent reinforcement learning with the idea of the time tunnel. *Energy, 153*, 977–987.
Xiong, R., Cao, J., & Yu, Q. (2018). Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle. *Applied Energy, 211*, 538–548.
Xu, H., Sun, H., Nikovski, D., Kitamura, S., Mori, K., & Hashimoto, H. (2019). Deep reinforcement learning for joint bidding and pricing of load serving entity. *IEEE Transactions on Smart Grid, 10*(6), 6366–6375.
Yan, J., He, H., Zhong, X., & Tang, Y. (2016). Q-learning-based vulnerability analysis of smart grid against sequential topology attacks. *IEEE Transactions on Information Forensics and Security, 12*(1), 200–210.
Yang, T., Huang, Z., Pen, H., & Zhang, Y. (2017). Optimal planning of communication system of CPS for distribution network. *Journal of Sensors* 2017.
Yang, T., Zhang, Y., Li, W., & Zomaya, A. Y. (2020). Decentralized networked load frequency control in interconnected power systems based onstochastic jump system theory. *IEEE Transactions on Smart Grid* Early access, https://ieeexplore.ieee. org/document/ 9023396.
Yang, T., Zhao, L., & Wang, C. (2019). Review on application of artificial intelligence in power system and integrated energy system. *Automation of Electric Power Systems, 43*(1), 2–14.
Ye, Y., Qiu, D., Li, J., & Strbac, G. (2019). Multi-period and multi-spatial equilibrium analysis in imperfect electricity markets: A novel multi-agent deep reinforcement learning approach. *IEEE Access, 7*, 130515–130529.
Ye, Y., Qiu, D., Sun, M., Papadaskalopoulos, D., & Strbac, G. (2020). Deep reinforcement learning for strategic bidding in electricity markets. *IEEE Transactions on Smart Grid, 11*(2), 1343–1355.
Yin, L., & Yu, T. (2018). Design of strong robust smart generation controller based on deep Q learning. *Electric Power Automation Equipment, 38*(5), 12–19.
Yin, L., Yu, T., & Zhou, L. (2018). Design of a novel smart generation controller based on deep Q learning for large-scale interconnected power system. *Journal of Energy Engineering, 144*(3), 04018033.
Yousefian, R., & Kamalasadan, S. (2015). Design and real-time implementation of optimal power system wide-area system-centric controller based on temporal difference learning. *IEEE Transactions on Industry Applications, 52*(1), 395–406.
Yu, T., & Zhen, W.-G. (2009). A reinforcement learning approach to power system stabilizer. *IEEE Power & Energy Society General Meeting*, 1–5 pp. 1-5.
Yu, T., Zhou, B., & Chan, K.-W. (2009). Q-learning based dynamic optimal CPS control methodology for interconnected power systems. *Proceedings of the CSEE, 29*(19), 13–19.
Yu, T., Zhou, B., Chan, K., Yuan, Y., Yang, B., & Wu, Q. (2012). R (λ) imitation learning for automatic generation control of interconnected power grids. *Automatica, 48*(9), 2130–2136.
Zarrabian, S., Belkacemi, R., & Babalola, A. A. (2016). Reinforcement learning approach for congestion management and cascading failure prevention with experimental application. *Electric Power Systems Research, 141*, 179–190.
Zeng, Q., Fang, J., Li, J., & Chen, Z. (2016). Steady-state analysis of the integrated natural gas and electric power system with bi-directional energy conversion. *Applied Energy, 184*, 1483–1492.
Zhang, X., Yu, T., Yang, B., & Cheng, L. (2017). Accelerating bio-inspired optimizer with transfer reinforcement learning for reactive power optimization. *Knowledge-Based Systems, 116*, 26–38.
Zhang, Z., Qiu, C., Zhang, D., Xu, S., & He, X. (2019). A coordinated control method for hybrid energy storage system in microgrid based on deep reinforcement learning. *Power System Technology, 43*(6), 1914–1921.
Zhao, H., Wang, Y., Guo, S., Zhao, M., & Zhang, C. (2016). Application of a gradient descent continuous actor-critic algorithm for double-side day-ahead electricity market modeling. *Energies, 9*(9), 725.
Zou, Y., Liu, T., Liu, D., & Sun, F. (2016). Reinforcement learning-based real-time energy management for a hybrid tracked vehicle. *Applied Energy, 171*, 372–382.