Applications of Reinforcement Learning in Deregulated Power Market:
A Comprehensive Review

Ziqing Zhu\textsuperscript{a}, Ze Hu\textsuperscript{a}, Ka Wing Chan\textsuperscript{a}, Siqi Bu\textsuperscript{a}, Bin Zhou\textsuperscript{b}, Shiwei Xia\textsuperscript{c}

\textsuperscript{a} Department of Electrical Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong SAR
\textsuperscript{b} College of Electrical and Information Engineering, Hunan University, Changsha, 410082, China
\textsuperscript{c} School of Electrical and Electronic Engineering, North China Electric Power University, Beijing, 102206, China

Abstract: The increasing penetration of renewable generations, along with the deregulation and marketization of power industry, promotes the transformation of power market operation paradigms. The optimal bidding strategy and dispatching methodology under these new paradigms are prioritized concerns for both market participants and power system operators, with obstacles of uncertain characteristics, computational efficiency, as well as requirements of hyperopic decision-making. To tackle these problems, the Reinforcement Learning (RL), as an emerging machine learning technique with advantages compared with conventional optimization tools, is playing an increasingly significant role in both academia and industry. This paper presents a comprehensive review of RL applications in deregulated power market operation including bidding and dispatching strategy optimization, based on more than 150 carefully selected literatures. For each application, apart from a paradigmatic summary of generalized methodology, in-depth discussions of applicability and obstacles while deploying RL techniques are also provided. Finally, some RL techniques that have great potentiality to be deployed in bidding and dispatching problems are recommended and discussed.

Key Words: Power market, Reinforcement Learning, bidding strategy, optimal dispatching

List of Abbreviations

| Abbreviation | Description |
|--------------|-------------|
| AC           | Actor-Critic algorithm |
| ASM          | Ancillary Service Market |
| CEAM         | Carbon Emission Auction Market |
| CNN          | Convolutional Neural Network |
| DA           | Day-Ahead |
| DDPG         | Deep Deterministic Policy Gradient |
| DP           | Dynamic Programming |
| DQN          | Deep Q-learning |
| DSO          | Distribution System Operator |
| EC           | Edge Computing |
| FL           | Federated Learning |
| GENCO        | Generation Company |
| HRL          | Hierarchical Reinforcement Learning |
| ISO          | Independent System Operator |
| LSTM         | Long Short Term Memory |
| MADDPG       | Multi-Agent Deep Deterministic Policy Gradient |
| MARL         | Multi-Agent Reinforcement Learning |
| MC           | Monte-Carlo method |
| MDP          | Markov Decision Process |
| MG           | Microgrid |
| NEP          | Nash Equilibrium Point |
| P2P          | Peer-to-Peer |
| RARL         | Risk-Averse Reinforcement Learning |
| RDG          | Renewable Distributed Generators |
| RL           | Reinforcement Learning |
| RT           | Real-Time |
| SPG          | Stochastic Policy Gradient |
| TD           | Temporal Difference method |
| VPP          | Virtual Power Plant |
1. Introduction

1.1 Bidding and Dispatching in Deregulated power markets

With the de-regulation of power industry since the 1990s, the modern power system with incorporation of energy market and multiple stakeholders has replaced the conventional vertical integrated paradigm in several European and American countries [1], while such marketization has played a significant role in the maximization of social welfare by determining the price of electricity via fair and competitive market transactions. In recent years, the problem of global climate change has become increasingly serious, and the growth of renewable energy, as well as the promotion of energy conservation and emission reduction, has become a major task of power industry. However, the increased penetration of renewable generations has brought great challenge to the power system real-time balancing because of the uncertainty of power output, as well as the deducted controllable capacity to be deployed for network balancing. Faced with this new challenge, the recent two decades have witnessed significant reformations of power market paradigms [2], in order to achieve a balance among the security of power system operation, economy of energy procurement, and environmental friendliness.

The features of the worldwide power market development can be summarized as follows: 1) The trading time-frame is shortened in order to tackle the rapid intermittence of high-penetrated renewable generation [3]; 2) More participants have been incorporated in the local power market [4], including the demand-side loads, virtual power plants (VPPs), and microgrids (MGs). 3) The coordinated market with several types of markets [5] are playing more significant roles, for instance, the coordination between electricity and auxiliary service market, and carbon emission trading auction market. Under these new paradigms, for both market participants’ and market operators’ concerns, the following two problems are under a pressing need to be investigated:

1) Optimal Bidding Strategy of Generation Companies (GENCOs):

The bidding procedure is one of the core components in the market transaction [6], in which GENCOs are required to submit their bidding price and available capacity, and the independent system operator (ISO) will decide the winners and capacity allocations to those winner GENCOs based on the market clearing rules. Considering previous experience is not applicable in new market paradigms, it is therefore a very natural concern of GENCOs regarding how to make their bidding decisions in order to maximize the revenues. However, it makes no sense to consider the individual GENCO’s bidding strategy regardless of other competitor GENCO’s, because whether the GENCO’s bid is successful or not is determined by its “ranking” among all the participated GENCOs, while the competitors will continuously modify their bidding strategies. Instead, to estimate the Nash equilibrium (NE) of the bidding game [7], i.e., the set of bidding strategies of all GENCOs that no one has the motivation to modify its bidding strategy, is a more applicable approach. Note that in most of literatures, the terminology of “to find the optimal bidding strategy” actually indicates “to compute the NE in the bidding game”. Hence, we hereinafter consider these two terminologies as equivalence in this paper.

The market equilibrium computation is the core issue for both GENCOs and ISOs [8], especially in the early stage of new market paradigm implementation. For GENCOs as market participants competing with each other for successful bids, they can precisely estimate the market clearing price and submit their bidding price slightly in order to mitigate the risk of bidding failure. ISO can detect potential behaviors that undermine fair market transactions, such as arbitrage and abuse of market power, by simulating all the possible scenarios of equilibrium before implementing new market rules, so that market regulations can be adjusted and the fairness of trading can be guaranteed based on the incentive compatibility principle [3].

2) Optimal Dispatching Strategy of the ISO

The economic dispatching problem (also called as “market clearing” and “energy management”) refers to the optimal allocation of power generation capacity for available generators in the whole network [9], in order to minimize the total generation cost, while satisfying the load demand and securing the network operation. Nowadays, the increasing amount of renewable generation and flexible load demand further complicates this problem because of the intermittence and non-predictableness of the total net load [10]. Such uncertainty will not only lead to the sub-optimality of solution in terms of economic efficiency, but also brings great challenge to the network real-time balancing. In addition, with the incorporation of new market paradigms [11], the dispatching problem needs further investigations in terms of: 1) the coordinated dispatching of multiple markets (for example, the generalized “power market” may include both electricity market and ancillary service market); 2) the impact of dispatching decisions on market operation, including the network congestion management, the mitigation of market power, etc.; and 3) the decentralized dispatching for autonomous GENCOs in the distribution network.

1.2 Applications of Reinforcement Learning on Bidding and Dispatching Problems

Given the aforementioned significance, the bidding and dispatching problems in the deregulated power market have been widely investigated in past literatures. For the bidding problem, a huge amount of works have focused on the risk – averse optimization of bidding strategies, including chance – constrained stochastic optimization [12,13], robust optimization [14,15], and distributional robust optimization [16,17], in which uncertainties of other GENCOs’ bidding strategy and the
market clearing price are considered. However, these methods simply neglect the modification of other GENCOs’ bidding strategies, and therefore cannot be used to compute the market equilibrium. In some literatures [18] – [20], the heuristic (group optimization) algorithms are adopted to simulate the procedure of dynamic bidding, but the obtained result is indeed the optimal Pareto frontier under the cooperative environment, in which agents are allowed to share their searching strategies with each other. However, in the competitive bidding game, the fully-distributed training without any communication among participated GENCOs is required, in order to facilitate the privacy protection and mitigate potential collusions which are harmful to the fairness of competition. For the dispatching problem, those risk-averse optimization techniques are also adopted in past researches [21] – [23]. However, the diversified operation characteristics (for example, on-and-off or ramping constraints for switches and generators), and the overwhelmingly large scale of expanded power networks, render the dispatching problem to be with huge dimensionality, non-convexity, and uncertainty, in which the adoption of conventional optimization tools will be very complicated in terms of mathematical derivations. The RL technique, as an agent-based method, enables the optimal decision searching of both the dispatching and bidding problem; meanwhile, the model-free characteristic of some RL algorithms removes the requirement of complicated mathematical modeling, and enables the agent to search the optimal decision in a more convenient manner by interacting with the environment.

Most importantly, both the bidding and dispatching problems imply the Markov Property, i.e., the decision at time slot \( t \) is influenced by the decision of the previous time slot \( t - 1 \). For instance, GENCOs will make their bidding decisions partially based on the market clearing result of the previous round of bidding; meanwhile, the ISO will also adjust their dispatching strategies based on the market and power network operation conditions, i.e., manipulation behaviors, abuse of market power, and network congestions. Considering the Markov property, the objective functions of these problems should not be the maximization of revenue in the single time slot; instead, it should be more farsighted – to maximize the total revenue in a long period of time. Hence, the bidding and dispatching problem can be considered as a dynamic programming with Markov property and uncertain environment, while these problems can be perfectly addressed by RL techniques. As a brief summary, advancements and feasibilities of RL compared with conventional optimization techniques are concluded in Table.1.

### Table 1. Comparison of Different Optimization Techniques

| Solutions                          | Computation Speed | Applicability in the Multi-Agent Environment | Applicability of Fully-distributed Training | Implication of Markov Property |
|------------------------------------|-------------------|---------------------------------------------|--------------------------------------------|--------------------------------|
| Mathematics-based Optimization     | Slow              | ✗                                           | ✗                                          | ✗                              |
| Stochastic Optimization            | Slow              | ✗                                           | ✗                                          | ✗                              |
| Distributionally Robust Optimization | Slow            | ✗                                           | ✗                                          | ✗                              |
| Heuristic Algorithms               | Fast              | ✓                                           | ✗                                          | ✗                              |
| Reinforcement Learning             | Fast              | ✓                                           | ✓                                          | ✓                              |

### 1.3 Bibliographical Status of Existing Reviews/Surveys

The recent two decades have witnessed a rapid development of RL techniques, which have catalyzed a sudden escalation of its applications on bidding and dispatching problems. The continuously increasing research activities have motivated the publication of some reviews/surveys, as briefly summarized in Table.2. These works are valuable materials for researchers and engineers in this discipline as guidance. However, according to the content listed in Table.2, limitations of previous reviews and surveys are summarized as follows:

- Currently there is no review/survey focusing on the operation of deregulated power markets, i.e., the simultaneous consideration of bidding and dispatching problems in a single paper. Some of them have a relatively large scope covering every potential application of RL in power systems including stability issues, fault diagnosis, etc., while some of them only discuss one single aspect of bidding and dispatching problems. Considering the bidding and dispatching are sequentially coupled problems in the market operation, it is necessary to simultaneously incorporate them into the discussion. In addition, the analysis in these reviews is not comprehensive, because the impact of market operation on the feasibility of RL applications and selection of RL algorithms are not taken into consideration.
- Some of them are focusing on very conventional RL algorithms, for example, Q-learning, Deep Q-Network (DQN), etc. Considering the extraordinarily rapid development of RL and the complexity of its application in power systems,
these algorithms are too old-fashioned to be adopted, while attention should be paid on those more state-of-the-art RL algorithms with significant improvements on the computational performance.

- Some of them are simply piling up existing research contributions, without in-depth and comprehensive discussions of the paradigmatic methodology of tackling different problems, pros and cons of different RL algorithms, and main obstacles of deploying RL in real-world situations, etc.
- The bibliographies cited in some of them are not appropriately selected. For instance, some cited literatures entitled with applications of RL algorithms in power systems do not contain any contribution on RL algorithm development or adoption; instead, they only adopted the methodology of dynamic programming and/or optimal control, without discussion of the Markov property and solutions using specified RL algorithms. In addition, due to the emerging trend of this concerned topic, a large quantity of very interesting and inspiring works have been published in the latest 2 years, but they are not cited in the existing reviews/surveys.
- The discussions of future trends in most of them are vague, without detailed elaborations in terms of research questions to be tackled and RL methods that can be adopted. Considering the application of RL in deregulated power markets is apparently lagging behind the development of RL techniques, there are many fancy RL algorithms developed in recent years that can be deployed to solve the bidding and dispatching problems; therefore, the applicability of these effective tools on operations of deregulated power markets should be discussed.

### Table 2. Summary of Existing Literature Reviews

| Ref. No. | Year | Covered Topics |
|----------|------|----------------|
| [24]     | 2021 | Frequency regulation, operational control, energy management (dispatching) |
| [25]     | 2021 | Building energy management systems, energy management (dispatching), ancill (bidding and dynamic pricing), vehicles and energy devices |
| [26]     | 2020 | Operational control, energy management (dispatching), power market (bidding and dynamic pricing), cyber security |
| [27]     | 2020 | Operational control, energy management (dispatching), power market (bidding and dynamic pricing), demand management |
| [28]     | 2019 | Energy management (dispatching), demand response, power market (bidding), operational control |
| [29]     | 2018 | Load/ power consumption forecasting, microgrids, demand response, defect/fault detection, cyber security, stability analysis |
| [30]     | 2017 | Operational control, energy management (dispatching), stability analysis |

#### 1.4 Main Contributions and Paper Structure

This paper aims to provide an up-to-date and comprehensive review in the operation of deregulated power markets, in terms of the bidding and dispatching problems, motivated by the aforementioned limitations of existing review papers. In this paper, the paradigmatic methodology of using RL techniques to tackle different problems in the deregulated power markets will be presented based on the cited existing literatures, along with in-depth discussions of potential obstacles and possible countermeasures. In addition, several most state-of-the-art RL techniques that have great potentiality to be deployed in future research are introduced, with discussions of their feasibilities of solving different problems. Specifically, contributions of this paper are outlined as follows:

- A comprehensive review, in the aspect of RL applications on bidding and dispatching problems, is presented in a systematic and paradigmatic manner. In this paper, the bidding and dispatching problems are classified into several sub-problems based on different application scenarios. For instance, the bidding problem is divided as the equilibrium analysis in the pool-based wholesale market and the bi-level pricing problem considering demand response; the dispatching problem is classified as the centralized and distributed dispatching. Each sub-problem is presented in the aspects of Markov Decision Process (MDP) model formulation, summary of applicable RL algorithms based on existing literatures, and feasibility of deploying different RL algorithms. Hence, this review can be served as a handbook for both researchers and engineers in this discipline to select and implement appropriate RL algorithms based on their requirements.

- In addition, for each sub-problem, a critical discussion on the applicability and potential obstacles of RL deployment in real-world implementations is provided, mainly in the aspects of the fairness of power market transactions, the economic efficiency of power market operation, and the security of power network operation. These aspects are top priorities in deregulated power markets, and must be taken consideration while justifying the applicability of new technologies. Meanwhile, potential solutions to tackle these problems are also briefly introduced.

- The cited literatures in this review are carefully selected, and the majority of them are published in the recent 5 years, in order to incorporate as many state-of-the-art RL techniques as possible. For the same reason, some papers with overly-simplified applications (i.e., the simple dynamic programming) are excluded from the bibliography of this
2. Fundamentals of Reinforcement Learning

In this section, a comprehensive review of RL fundamentals is provided, covering all necessary concepts and algorithms that will be further employed while elaborating RL applications on deregulated power markets in subsequent sections. Firstly, the Markov Decision Process (MDP) as the most simplified formulation of RL is introduced, along with the methodology of dynamic programming (DP) as its solution, under the assumption of full-observability of MDP elements, i.e., the “model-based” RL. Then, the MDP is further extended to the condition with unknown MDP model, i.e., “model-free”, which is addressed by the value-based RL (including Monte Carlo (MC) and Temporal Difference (TD) methods) and policy-search-based RL (including stochastic and deterministic policy gradient methods). In addition, several widely-adopted algorithms developed based on these methodologies are also briefly introduced, including Q-learning, Deep Q-learning (DQN), Trust Region/Proximal Policy Optimization (TRPO/PPO), and Deep Deterministic Policy Gradient (DDPG) with the framework of actor-critic (AC), etc. Finally, several state-of-the-art multi-agent RL algorithms are supplemented, which are frequently deployed especially in literatures of the recent two years because of its consistency with the trend of deregulation and decentralization of power systems. Fig.1. illustrates the general classification of aforementioned RL algorithms, and the following elaborations will also follow the framework as depicted.

Fig.1 Classification of Reinforcement Learning Algorithms

2.1 Markov Decision Process (MDP)

2.1.1 Properties and Elements of MDP

To start with, we will firstly introduce the concept of Markov Property as a foundation of Markov Process (MP) and MDP. The Markov Property is numerically defined as [31]:

\[ P[s_{t+1} | s_t] = P[s_{t+1} | s_t, s_{t-1}, ..., s_1] \]

which refers to the conditional probability of state \( s_{t+1} \) occurring given that \( s_t \) has already occurred, being independent of the previous states from \( s_{t-1} \) and beforehand. Intuitively, the MP is defined as the state sequence with such a property. The MP is always formulated as a two-tuple \((S, P)\), where \( S \) is a set of states with Markov Property, and \( P \) denotes transition probabilities among states. By introducing a smart agent that can take different actions in different states, the MP can be extended to the MDP formulated as a tuple \((S, A, R, P)\) [32] with actions and rewards, in which

- \( S \) is a set of finite states.
• $A$ is a set of finite actions that can be taken by the agent in the corresponding state.
• $P$ represents the transition probability, denoted as: $P: S \times A \rightarrow (S)$, which is a probability distribution over the set $S$ (i.e., it assigns probabilities to states).
• $R(S, A)$ indicates the reward of selecting action $a$ in the previous state $s$.

2.1.2 Bellman Equation

Once the MDP is observed (we assume the MDP is fully-observable in this subsection, while the condition of partially-observable MDP will be discussed later), the aim of agents is to obtain the optimal policy $\pi^*$, which refers to the sequential decisions [33] mapping states to actions, in order to maximize the total rewards in MDP. However the “performance” of a given policy cannot be simply evaluated by the immediate reward after the action is determined because of the consideration of long-term benefits. Instead, the state-value function as formulated in (2), which indicates the so-called “expected accumulative reward”, is always used to evaluate the performance of implementing the policy $\pi$ at the state $s$ by interacting with the environment till the termination of the entire episode. To update the policy by choosing actions that lead to states with more rewards may be a solution, but the mapping from actions to states is also with uncertainty (see the definition of transition probability). Intuitively, the action-value function is introduced as shown in (2), which implies the expected accumulated reward by executing the policy $\pi$ after taking an action $a$ at the state $s$. Note that the discount factor $\gamma$ is intended to disqualify the reward indicating the uncertainty in the future.

\[
V_\pi (s_t) = E_\pi \left[ \sum_{k=0}^{\infty} \gamma^k R_{k+1} \right] \\
Q_\pi (s_t, a_t) = E_\pi \left[ \sum_{k=0}^{\infty} \gamma^k R_{k+1} \right]
\]

(3) indicates that the action-value constitutes two parts: the immediate reward, and the sum of possible state-values at $s_{t+1}$ weighted by their probabilities. Inspired by (3), the “optimal” policy can be induced by taking the action with maximum expected reward in an iterative manner once the action-value function is available. Furthermore, by substituting $G_{t+1} = V_\pi (s_{t+1})$ (see the definition of $G_t$), the iterative expression of $V_\pi (s_t)$, which is the so-called Bellman Equation [34], can be written as

\[
V_\pi (s_t) = E_\pi \left[ R_{t+1} + \gamma V_\pi (s_{t+1}) \right] \\
Q_\pi (s_t, a_t) = E_\pi \left[ R_{t+1} + \gamma Q_\pi (s_{t+1}, a_{t+1}) \right]
\]

By definition, the optimal state-value function and action-value function are the maximum values of $V_\pi (s_t)$ and $Q_\pi (s_t, a_t)$, i.e., $V^*(s_t) = \max_{\pi} V_\pi (s_t)$ and $Q^*(s_t, a_t) = \max_{\pi} Q_\pi (s_t, a_t)$. Then, the Bellman Equation of the optimal state-value function and action-value function can be formulated as follows [34], where $P_{s_t \rightarrow s_{t+1}}^a$ is the transition probability from $s_t$ to $s_{t+1}$ after taking an action $a_t$.

\[
V^*(s_t) = \max_a R(s_t, a_t) + \gamma \sum_{s_{t+1}} P_{s_t \rightarrow s_{t+1}}^a \max_{a'} V^*(s_{t+1}) \\
Q^*(s_t, a_t) = R(s_t, a_t) + \gamma \sum_{s_{t+1}} P_{s_t \rightarrow s_{t+1}}^a \max_{a'} Q^*(s_{t+1}, a_{t+1})
\]

With (7) in hand, the optimal policy $\pi^*$ can be intuitively derived by maximizing the action-value $Q^*(s_t, a_t)$ as formulated in (8), while such procedure is implemented with several useful techniques including DP, value-based RL, policy-searching-based RL, etc., as indicated in Fig.1 and elaborated in the subsequent subsections.

\[
\pi^*(a_t | s_t) = \begin{cases} 1, & \text{if } a = \arg \max_{a'} Q^*(s_t, a_t) \\ 0, & \text{otherwise} \end{cases}
\]

2.2 Model-based Dynamic Programming (DP)

The DP is the most straightforward and simple method to solve the MDP with full observability (i.e., “model-based”) and find the optimal policy. As the name implies, “dynamic” indicates the sequential decision-making, and “programming” refers to “optimization”. In general, the DP constitutes two steps: policy evaluation and policy improvement, in which the former refers to the evaluation of the value of a given policy, and the latter refers to updating the policy function to improve the evaluated “values”. The evaluation of value of any given policy is based on the following equation [34]:

\[
V_\pi (a_t) = \sum_{s_{t+1}} \pi (a_t | s_t) \left[ R(s_{t+1}, a_{t+1}) + \gamma \sum_{s_{t+1}} P_{s_{t+1} \rightarrow s_{t+2}}^a \max_{a'} V^*(s_{t+2}) \right]
\]

To be specific, the value function of a given policy in the current state is expressed by the value function in the next state and other known parameters $R, S, A, \gamma$. The method is called “bootstrapping” because the calculation of value function in the
current state requires its value in the next state, which is updated by the technique of Gauss-Seidel iteration [34]. After the policy evaluation, the greedy strategy [35] is adopted to improve the policy as shown in (10), by means of selecting actions with the maximum sum of Q-values in each state. The aim of greedy strategy is to update the policy for the next iteration \( i + 1 \), while the relationship between the “evaluated value” and the “Q-value” is indicated in (9):

\[
\pi^{i+1}(s_t) \in \arg \max_a Q_{\pi^i}(s_t, a_t)
\]

(10)

It is important to emphasize that, the DP is only feasible to solve the Bellman equation subject to the low dimensionality of state-action space in the MDP to avoid the “curse of dimensionality” [36], as well as the full observability of state transition probability. However, for real-world implementations, these two conditions are very hard to be fulfilled. To address this problem, the model-free RL, which constitutes the value-based and policy-search-based approach, has been developed with improved computational performance in recent years. This branch of RL technique is reviewed in the following subsections.

2.3 Value-based RL Method

From this subsection and thereafter, we will focus on the model-free RL, i.e., the MDP is not fully-observable. The generalized methodology resembles that of model-based RL, constituting two main steps: policy evaluation in a bootstrapping manner and policy improvement using the greedy strategy. However, recall (9) which contains the transition probability, and it is obvious that the approach introduced previously cannot be directly adopted herein. To tackle this issue, two well-known techniques have been developed to estimate the value function, including the Monte-Carlo method [37] and Temporal-Difference method [38], which are classified into the value-based RL, differentiating from policy-searching-based RL to be introduced in Section 2.4.

2.3.1 Monte-Carlo-based RL Method

As mentioned above, the Monte Carlo (MC) method is deployed to estimate the value function when the MDP formulation is not available, and the general methodology is to compute the average value of sampled rewards. Once the value function is estimated, those policy improvement methods mentioned in Section 2.2 can be directly deployed to pursue the optimal policy. The detailed pseudocode of MC is available in [37], followed by some important remarks in order to explain how the dataset is sampled via interacting with the environment, how to compute the “average” value, and how to implement the policy improvement.

Remark 1. (Exploration and Exploitation) In order to obtain the optimal policy function, the agent needs to address the dilemma of “ensuring the obtained policy being not worse than the current optimal policy” (exploitation) and “explore the environment as much as possible to search for potential better policy” (exploration). Hence, for the agent making the decision of its action in the next state, despite the objective of maximizing the Q-value, the access to each state-action pair should also be guaranteed. For instance, a well-known and frequently-adopted strategy aligning with the requirement of tackling such dilemma is the “\( \varepsilon \)-greedy” [39], in which the parameter \( \varepsilon \) indicates the probability of “exploration” (i.e., to take a random action in the next step), and the probability of “exploitation” (i.e., to take the action with the maximum Q-value) is therefore \( 1 - \varepsilon \).

Remark 2. (On-policy and Off-policy) Note that the aforementioned \( \varepsilon \)-greedy refers to the policy for interacting with the environment in the training procedure, i.e., the so-called “behavior policy” [40], which is differentiated from the “target policy” referring to the policy for actual implementation after training. According to whether the behavior policy \( \pi \) and target policy \( \mu \) are the same, the Monte Carlo method is further divided into two categories: on-policy or off-policy. On-policy implies the same policy for exploration and exploitation, while off-policy avert the potential local optimality in the on-policy method by differentiating the behavior and target policy, while adopting the important sampling technique to evaluate and update the target policy using the dataset sampled by the behavior policy.

2.3.2 Temporal-Difference-based RL Method

As indicated above, the estimation of value function in MC requires a large amount of sampled data, and therefore cannot be completed before the end of one entire episode, which will result in an agonizingly slow computational speed. In order to accelerate the computation, an intuitive idea is to adopt the bootstrapping method mentioned in DP, in which the value function of current state is estimated based on its value of the next iteration. To be concrete, the value function update of TD is shown in (11), where \( V(s_{t+1}) \) is the value of the next state estimated from the experiences, and \( \beta \) indicates the learning rate of the algorithm. Regarding the reward and discount for the future, \( R_{t+1} + \gamma V(s_{t+1}) \) and \( R_{t+1} + \gamma V(s_{t+1}) - V(s_t) \) indicate the target of value and TD error standing for the error between current value and target respectively [34]:

\[
V(s_t) \leftarrow V(s_t) + \beta \left( R_{t+1} + \gamma V(s_{t+1}) - V(s_t) \right)
\]

(11)

Hence, the TD method can be considered as the combination of DP and MC, inheriting both of their advantages; for instance, the sampling in MC enables the training without the observability of transition probability, while the bootstrapping method
effectively mitigates the computational burden of reduplicative sampling. Several well-known TD-based RL algorithms include SARSA and Q-Learning, which are widely used in literatures cited in this review, while the former adopts the on-policy method and the latter inherits the methodology of off-policy. The detailed pseudo code of these two algorithms is presented in [41] and [42].

2.3.3 Function-approximation-based RL Method

Aforementioned methods, including DP, MC and TD, are introduced to handle problems with discrete and limited state-action spaces by means of value evaluation and policy improvement. To be noticed, the value function in these algorithms is developed in the form of Q-tables with the index of state-action pairs, in which corresponding Q-values are updated within the policy improvement. These algorithms are therefore classified as tabular RL, fitting to problems with discrete and low-dimension of state-action space, to mitigate the “curse of dimensionality”, which refers to the exponential expansion of state-action space with the increase of dimension of states and actions. Such phenomenon will not only lead to the overwhelming computational burden, but also result in the obtained policy function trapping into the local optimum. Facing such problems, inspired by the methodology of supervised learning, the so-called function approximation method is proposed along with the parameterized value-function \( V(s_t, \theta) \) relating to state \( S \), action \( A \) and a set of parameters \( \theta \). This method exhibits a more favorable computational performance, by approximating such value-function by updating parameter \( \theta \) iteratively through (12) [43], where \( \hat{Q}(s_t, a_t, \theta) \) denotes the approximated action-value function.

\[
\arg\min_{\theta} \left( Q(s_t, a_t) - \hat{Q}(s_t, a_t, \theta) \right)^2
\]  

(12)

Two representative algorithms are introduced in this subsection, including Deep Q-learning (DQN) [44] and its modified version Double Deep Q-learning (DDQN) [45], which are frequently adopted for applications on topics that are concerned in deregulated power markets. DQN is a well-known function approximation RL algorithm modified from Q-learning with the following threefold advancements. Firstly, the deep convolutional neural network (CNN) is adopted to approximate the value function. Secondly, the trick of experience replay is applied to DQN by restoring the sampled dataset in the replay buffer and drawing them randomly while implementing the training. This trick is intended to remove the probabilistic correlations among sampled data, in order to fulfill the requirement of independent and identically distributed data while training the neural network. At last, instead of using the real-time updated Q-value produced by the original Q-network, an independent TD-network is established in DQN so as to calculate the TD error, which will be updated Q-value’s update, consequently reducing the instability of network training caused by the same parameters of the Q-network and TD-network. However, the DQN also inherits the limitation of Q-learning, in terms of the over-estimation of the Q-value. Such over-estimation will negatively contribute to the accuracy of estimated Q-values, rendering the sub-optimal state-action pair being considered as the optimal solution by mistake. To address this problem, the DDQN method is proposed, in which two different Q-values are used in the procedure of selecting the current action and evaluating the current state-action pair. The detailed pseudo code of DQN is presented in [44, 45].

2.4 Policy-search-based RL Method

The aforementioned value-based RL methods are intended to solve MDP with discrete action space and discrete/continuous state space, by updating the policy based on the value function. Whereas the DQN utilizing neural networks like CNN to optimize the parameterized value function directly, it exhibits a favorable performance in terms of the computational efficiency. However, these algorithms are not capable of solving problems with continuous action-space, in which the value of argmax \( Q(s_t, a_t) \) in (10) cannot be numerically obtained. In this subsection, the policy-search-based RL algorithms will be introduced to tackle this problem. By adopting these algorithms, the parameterization of policy functions (rather than value functions) not only enables the consideration of continuous action space, but also improves the learning efficient and capability of convergence considerably. The policy-search-based RL can be divided into model-free and model-based manner, while the former constitutes the stochastic and deterministic policy gradient algorithms. In the follows, two selected algorithms including the TRPO/PPO [46,47] and DDPG [48] are introduced, as typical and frequently-used representatives of stochastic and deterministic policy gradient respectively.

2.4.1 Stochastic Policy Gradient (SPG)

Before elaborating the detailed algorithms, it is necessary to develop the concept of policy gradient in RL. Consider a MDP problem with a trajectory of state-action pair \( \pi; s_\tau, a_\tau, \ldots, s_H, a_H \), and \( R(\tau) = \sum_{t=0}^{H} R(s_t, a_t) \) indicates the reward and \( P(\tau; \theta) \) indicates the probability distribution of all trajectories in this stochastic policy, where \( \theta \) are parameters of the policy. To adopt RL algorithms to find the optimal policy of this MDP, the “objective” can be intuitively written as follows [46]:

\[
J(\theta) = E\left[ \sum_{t=0}^{H} R(s_t, a_t); \pi_\theta \right] = \sum_\tau P(\tau; \theta) R(\tau)
\]

(13)
max \ J(\theta) = \max \ \theta \, \sum P(\tau; \theta) R(\tau) \quad (14) 

where \ J(\theta) \ is \ the \ expected \ accumulative \ rewards. \ In \ other \ words, \ the \ goal \ of \ the \ adopted \ RL \ algorithm \ is \ to \ find \ the \ maximized \ expected \ accumulative \ reward, \ i.e., \ \max J(\theta), \ by \ calculating \ the \ optimal \ parameter \ \theta \ with \ the \ employment \ of \ optimization \ methods, \ such \ as \ steepest \ descent, \ Newton \ method, \ and \ interior \ point \ method. \ Among \ these \ methods, \ the \ steepest \ descent, \ also \ known \ as \ “policy \ gradient”, \ is \ the \ most \ straightforward \ way \ to \ solve \ the \ aforementioned \ problem, \ by \ calculating \ the \ value \ of \ \nabla_\theta J(\theta) \ in \ the \ steepest \ descent \ formulation:

\[ \theta_{\text{new}} = \theta_{\text{old}} + \alpha \nabla_\theta J(\theta) \quad (15) \]

where \( \alpha \) \ indicates \ the \ step \ size. \ Now \ take \ the \ derivative \ of \ objective \ function \ (\theta), \ as \ formulated \ in \ (16) [46]:

\[ \nabla_\theta J(\theta) = \sum P(\tau; \theta) \nabla_\theta \log P(\tau; \theta) R(\tau) \quad (16) \]

where \ the \ policy \ gradient \ can \ be \ considered \ as \ the \ expectation \ of \ \nabla_\theta \log P(\tau; \theta) R(\tau), \ which \ is \ estimated \ by \ the \ sampled \ \( m \) \ trajectories \ via \ interacting \ with \ the \ environment \ under \ the \ policy \ \theta, \ as \ re-formulated \ in \ (17):

\[ \nabla_\theta J(\theta) \approx \hat{g} = \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \log P(\tau_i; \theta) R(\tau_i) \quad (17) \]

in which \( \hat{g} \) \ indicates \ the \ policy \ gradient, \( m \) \ indicates \ the \ number \ of \ sampling \ trajectories, \ and \ the \ term \ \nabla_\theta \log P(\tau; \theta) \ indicates \ the \ logarithmic \ probability \ of \ sampling \ trajectory \ \tau \ under \ the \ policy \ \theta, \ while \ the \ logarithmic \ operator \ is \ intended \ to \ accelerate \ the \ convergence \ by \ mapping \ the \ probability \ from \ [0,1] \ to \ [-\infty, 0]. \ It \ is \ obvious \ that \ the \ “direction” \ and \ step \ size \ of \ policy \ update \ are \ dependent \ on \ the \ polarity \ and \ value \ of \ the \ total \ reward \ \( R(\tau) \) \ in \ the \ sampled \ trajectory. \ Therefore, \ it \ is \ intuitive \ that \ the \ policy \ gradient \ is \ intended \ to \ increase \ the \ probability \ of \ sampling \ trajectories \ with \ higher \ rewards.

It \ is \ important \ to \ note \ that, \ as \ indicated \ in \ (17), \ the \ determination \ of \ step \ size \ is \ the \ one \ of \ the \ most \ challenging \ problems \ in \ the \ SPG-based \ RL \ algorithm, \ because \ the \ inappropriate \ manipulation \ will \ result \ in \ the \ solution \ trapping \ into \ the \ local \ optimum. \ The \ Trust \ Region \ Policy \ Optimization \ (TRPO) \ algorithm \ is \ therefore \ proposed \ to \ tackle \ this \ problem, \ by \ re-formulating \ the \ objective \ performance \ (18) \ in \ the \ form \ of \ non-decreased \ function:

\[ \eta(\pi) = E_{t[s]} \left[ \sum_{t=0}^{\infty} \gamma^t (R(s_t)) \right] \quad (18) \]

\[ \eta(\bar{\pi}) = \eta(\pi) + E_{s_0, a_0, ..., s_T, a_T} \left[ \sum_{t=0}^{\infty} \gamma^t A_z(s_t, a_t) \right] \quad (19) \]

where \( \gamma \) \ indicates \ a \ set \ of \ state-action \ pairs \ \( s_0, a_0, ..., s_T, a_T \), \ and \( \pi, \bar{\pi} \) \ denote \ the \ current \ and \ updated \ policy \ function \ respectively. \ In \ particular, \ the \ term \ of \ \( A_z(s_t, a_t) \) \ denotes \ the \ so-called \ “advantage \ function”, \ which \ indicates \ the \ difference \ between \ the \ Q-value \ of \ the \ optimal \ state-action \ pair \ and \ the \ value \ function \ of \ the \ current \ state, \ as \ shown \ below:

\[ A_z(s_t, a_t) = Q_z(s_t, a_t) - V_z(s_t) = E_{s_{t+1} \sim P(s_{t+1} | s_t, a_t)} \left[ R(s_t) + \gamma V_{\bar{\pi}}(s_{t+1}) - V_{\bar{\pi}}(s_t) \right] \quad (20) \]

where \( Q_{\pi}(s_t, a_t) \) \ and \( V_{\pi}(s_t) \) \ are \ state-action \ value \ function \ and \ value \ function \ respectively, \ and \ the \ former \ is \ definitely \ larger \ than \ the \ latter. \ Hence, \ the \ non-negativity \ of \ this \ advantage \ function \ can \ ensure \ the \ policy \ function \ being \ continuously \ and \ monotonically \ updated \ towards \ the \ optimality, \ which \ is \ the \ key \ insight \ of \ TRPO \ algorithm. \ In \ addition, \ several \ further \ propositions \ are \ made \ to \ simplify \ the \ algorithm \ implementation, \ and \ the \ simplified \ version \ is \ so-called \ PPO \ algorithm \ as \ introduced \ in \ [47].

2.4.2 Deep Deterministic Policy Gradient (DDPG)

Although the SPG-based RL has been widely deployed, it still inherits the following limitations: firstly, the output of these algorithms is a policy function in the form of probabilistic distribution, which means the action is still not available before sampling in the state-action space with a high dimensionality, while such sampling will result in non-necessary computational burdens; secondly, the estimation of policy gradient using MC method, as indicated in (16), also requires such sampling that aggravates the computational resource consumption. To address these problems, the DDPG algorithm is proposed in [48] that combines the methodology of deterministic policy gradient (DPG) and actor-critic (AC). The framework of DDPG is depicted in Fig.2, with the following definitions to be firstly clarified:

- “Policy-network” and “Q-network”: The DDPG adopts the framework of AC, in which the “Policy-network” (actor) and “Q-network” (critic) plays the role of policy improvement and policy evaluation in two separate neural networks respectively. The “Policy-network” (actor) receives the sampled data obtained by interacting with the environment, and updates its policy function to optimize the performance objective, which is the expectation of Q-value of the current policy, while the required Q-value is subsequently estimated by the “Q-network” (critic). The “Q-network” is updated using the same methodology as the DQN.
The simultaneous adoption of Q-network’s parameter in updating the Policy-network and Q-network itself will result in the non-stationarity of training. To tackle this problem, two sets of Policy-network and Q-network are developed, namely the “Online-network” and “Target-network”, while the Online-network will be updated for each mini-batch of training, and the Target-network will be updated using the so-called soft update method based on the updated Online-network. The subsequent update of Policy-network will use the parameter of the Target-network with more stationarity instead of the online-network.

In order to facilitate the exploration, the DDPG algorithm is trained in an off-policy manner, in which the target policy and behavior policy are differentiated, as aforementioned in Section 2.3. For instance, the behavior policy is obtained by randomization of the trained target policy via the Uhlenbeck-Ornstein (UO) noise, and the trajectories of dataset are sampled by the behavior policy interacting with the environment.

**WoLF-PHC Algorithm** [49]: This algorithm can be considered as an extension of Q-learning to multi-agent environment, with the following advancements to be clarified. Firstly, as the name implies, the step of policy updating is variable based on the “win” or “lose” status of the state-action pair being evaluated. The determination of win or lose is based on the comparison of Q-value of the current policy and the so-called average policy, which represents the average Q-value that could be attained. Specifically, the policy function of the kth agent is updated in the following method:

$$\pi_k(s_t, a_t) = \pi_k(s_t, a_t) + \Delta_{sa}$$

where

$$\Delta_{sa} = \begin{cases} -\Delta_{sa}, & \text{if } a_t \neq \arg \max Q_k(s_t, a_t) \\ \Delta_{sa}, & \text{o.w.} \end{cases}$$

The updating coefficient is computed by
\[ \Delta_{\mu_a} = \min \left( \pi_t(s_t^{k,h}, a_t^{k,m}), \frac{\delta}{|a_t^{k,m}| - 1} \right) \]  

where \( \delta \) is the ancillary hyper-parameter, \( h \) and \( m \) denote the dimensionality of state and action space of the \( k \)th agent. The “win” state indicates the current policy has a higher Q-value than the average policy, and the agent will slowly increase the probability of selecting this policy with a small updating coefficient as a hyper-parameter. On the other hand, the “lose” state indicates the current policy has a lower Q-value than the average policy, and the agent will decrease the probability of selecting this policy with a large updating coefficient. By adopting this method, the agent could possibly accelerate the speed of convergence by mitigating the “bad policy” with lower Q-value. Each agent will only update its own Q-table, without requirements of availability of other agents’ Q-tables (policies). Hence, this algorithm is feasible to be adopted in the competitive environment with strict requirements of communication restrictions and privacy protection.

- **MADDPG Algorithm** [50]: This algorithm can be considered as an extension of DDPG to multi-agent environment, with the following problems to be tackled. Firstly, as proved in [50], the probability of sampling the “true” gradient which leads to the optimal policy will be exponentially decreased with the increase of number of agents. Secondly, as for the policy-search-based RL, the non-observability of other agents’ policies will render the convergence intractable. To tackle these two problems, the following tricks are proposed and effectively improve the computational performance:

1) **Centralized training, distributed implementation**: In this proposed algorithm, each agent will optimize its own policy function by updating the actor network in a distributed manner; however, these agents will use the common critic network with observation of global information for policy evaluation, while the availability of global information will enable the environment to be considered as stationary. To be concrete, consider the expected cumulative reward (ECR) \( J_\beta(\mu^k) = \mathbb{E}_{s_0 \sim P_\beta} \left[ Q^k(s_0, a_0^k, \mu^k(s_0, a_0^k)) | a_0^k \sim \mu^k(s_0) \right] \), the deterministic policy gradient can be derived as (24):

\[ \Delta_{\theta^k} J_\beta(\mu^k) = E_{s_t^k, a_t^k \sim D} \left( \nabla \theta^k Q^k(s_t^k, a_t^k, \mu^k) - \nabla \theta^k Q^k(s_t^k, a_t^k, \mu^k) \right) \]

where \( D \) denotes the experience replay buffer (ERB) containing the elements of \( \{s_t^k, a_t^k, R_t^k, k \in K\} \), and \( \mu^k \) denotes the parameterized policy function of the \( k \)th agent, differentiating from that of discrete policy \( \pi(s_t^k, a_t^k) \) in the tabular form. The dataset in the ERB will be transmitted to the critic network for centralized updating, using the back-propagation approach by minimizing the loss function \( L^k = \frac{1}{n} \sum (y_t^k - Q_t^k(s_t^k, a_t^k))^2 \), where \( y_t^k = R_t^k(s_t^k, a_t^k) + y Q^k(s_{t+1}^k, a_{t+1}^k, \mu^k) \).

2) **Estimation of other agents’ policies**: In order to relax the requirement of availability of other agents’ policy while updating the common critic network, another important trick is to replace this term by the estimation made by each agent, using the loss function with the entropy term, as formulated in (25):

\[ \chi(\theta^k) = -E_{a_t^k \sim \pi^k} \left( \log \mu_{\theta^k}^a(\pi^k | s_t^k) + \lambda H(\mu_{\theta^k}^a) \right) \]

where \( \mu_{\theta^k}^a \) denotes the policy function of other agents except for the \( k \)th agent. By minimizing the loss function, each agent will be enabled to estimate other agents’ policy functions without communications.

3) **Policy ensemble**: Due to the continuous update of other agent’s policy function, the obtained “optimal” policy is always over-fitting. To tackle this problem, each agent is designated to update a “set” of policy consisting of several sub-policies, and the objective performance is therefore the maximization of total ECR in this policy set. Hence, the robustness of this algorithm is effectively improved.

### 3. Applications of Reinforcement Learning in Deregulated Power Market Operation

#### 3.1 Bidding and Pricing Strategy Optimization in Power market

As mentioned in Section 1.1, the dynamic optimization of bidding strategy is equivalent to the computation of NE in the competitive bidding game, which is the key concern for both GENCOs and ISO. In this section, the application of RL on competitive bidding game, as well as its extension – the dynamic pricing considering demand response, will be elaborated separately in two subsections. For each subsection, we will start from the description of the generalized market framework and regulations, followed by the paradigmatic MDP formulation and summary of applicable RL algorithms deployed in existing literatures; meanwhile, some selected significant modifications of existing RL methods will be emphasized. Finally, an in-depth discussion will be presented in terms of applicability of different RL algorithms, and main obstacles of RL application in real-world implementations with potential solutions.
3.1.1 Dynamic Bidding Strategy Optimization using Multi-Agent Reinforcement Learning

We herein consider two typical frameworks of power market architecture: the centralized pool-based market, and the distributed peer-to-peer (P2P) market. The generalized market framework and time-frame are illustrated in Fig.3. For the former paradigm [51], in a single trading time-slot, the demand-side customers will firstly submit their load demands to the ISO who serves as the coordinator responsible for the market operation and management. The ISO will then broadcast the total load demand to the participated GENCOs, who will thereafter submit their individual available capacity and corresponding bidding price to the ISO for market clearing. Note that the market clearing is considered as an optimization problem, in which the objective function is to minimize the cost of generation (purchasing the energy from GENCOs), while several constraints for security operation must be satisfied. In the P2P market paradigm [52], the trading is implemented in a distributed and bilateral manner, while a coordinator is always necessary to release the information of bids and offers, and to conduct the match-making [53] using the merit-order with the objective of maximization of total social welfare.

Remark 3. In most of real-world power markets, the majority of energy transactions are completed in the day-ahead (DA) market, while the real-time (RT) market is always served as a complementation for the difference of net load prediction between DA and RT. Hence, most of existing literatures in terms of the bidding problem focus on the DA market, i.e., the Phase 2 and Phase 3 in Fig.3.

Remark 4. Fig.3 is only served as a paradigmatic, generalized and simplified illustration, indicating the time-frame, the procedure of energy transaction and procurement in power markets. It is assumed that both DA and RT markets are hourly-based, while in some countries the time-slot is 30 minutes or 15 minutes. Also, the parameters \(a, b, c\) in this figure indicating the time-slot of each phase are also dependent on specific market designs.

As discussed previously, to implement RL algorithms for bidding procedure simulation, the MDP should be firstly developed in (26) as a generalized form of bi-level dynamic programming, as shown in Fig.4, and the elements of state and action sets are detailed in Table.3. The upper-level denotes the market clearing procedure of ISO, subject to the network security constraints, as well as the bidding price and available capacity claimed by the GENCOs. The lower-level refers to the dynamic bidding strategy optimization of participated GENCOs, subject to operational characteristics and limitations of different types of generators. Note that although both pool-based and P2P trading follow the same generalized bi-level framework, their MDP elements are slightly different, which are separately listed in Table.3. Also, the annotated description of objectives and constraints for both ISO and GENCOs are summarized in Table.4, with detailed numerical expressions available in [54]–[89].

Fig.3. Time-frame of Generalized Power market
Upper Level:

\[
\max_{\mu_{\text{ISO}}(A_{\text{ISO}} | S_{\text{ISO}})} E \left[ \sum_{k=0}^{T-t} y_t R_{\text{ISO}}^{t+k} | S_{\text{ISO}}^{t+k} \right]
\]

s.t.

\[
A_{\text{ISO}}^{t} \in A_{\text{ISO}}
\]

\[
\hat{S}_{\text{ISO}}^{t} = \arg \max_{\mu_{\text{Gen}}(\alpha_{\text{Gen}} | S_{\text{Gen}})} E \left[ \sum_{k=0}^{T-t} y_t R_{\text{Gen}}^{t+k} | \mu_{\text{Gen}} \sim S_{\text{Gen}}^{t} \right] \in S_{\text{ISO}}
\]

Lower Level:

\[
\max_{\mu_{\text{Gen}}(\alpha_{\text{Gen}} | S_{\text{Gen}})} E \left[ \sum_{k=0}^{T-t} y_t R_{\text{Gen}}^{t+k} | \mu_{\text{Gen}} \sim S_{\text{Gen}}^{t} \right]
\]

s.t.

\[
A_{\text{ISO}}^{t} \in A_{\text{ISO}}
\]

Table 3. MDP Elements of Bidding Strategy Optimization Problem

| Market          | Elements | Details and Descriptions                                                                 |
|-----------------|----------|-----------------------------------------------------------------------------------------|
| Pool-based      | \(A_{\text{ISO}}^{t}\) | The computed market clearing price                                                      |
|                 | \(p_{\text{Gen},allo}^{t}\) | The amount of allocated generation to the \(i\)th participated GENCO                    |
| \(S_{\text{ISO}}^{t}\) | \(p_{\text{Gen},i}^{t}\) | The amount of available capacity claimed by the \(i\) th participated GENCO             |
|                 | \(\lambda_{\text{Gen},i}^{t}\) | The bidding price claimed by the \(i\)th participated GENCO                             |
| \(A_{\text{Gen}}^{t}\) | \(p_{\text{Gen},i}^{t}\) | The amount of available capacity claimed by the \(i\) th participated GENCO             |
|                 | \(\lambda_{\text{Gen},i}^{t}\) | The bidding price claimed by the \(i\)th participated GENCO                             |
| \(S_{\text{Gen}}^{t}\) | \(p_{\text{disp}}^{t}\) | The amount of dispatchable generators’ capacity                                         |
|                 | \(L_{\text{load}}^{t}\) | The amount of announced total load demand                                              |
| P2P Market      | \(A_{\text{ISO}}^{t}\) | The match-making result of P2P trading of the \(i\)th participated GENCO               |
| \(S_{\text{ISO}}^{t}\) | \(p_{\text{Gen},i}^{t}\) | The amount of available/insufficient capacity claimed by the \(i\) th participated GENCO |
|                 | \(\lambda_{\text{Gen},i}^{t}\) | The bidding/offering price claimed by the \(i\)th participated GENCO                   |
| \(A_{\text{Gen}}^{t}\) | \(p_{\text{Gen},i}^{t}\) | The amount of available/insufficient capacity claimed by the \(i\) th participated GENCO |
|                 | \(\lambda_{\text{Gen},i}^{t}\) | The bidding/offering price claimed by the \(i\)th participated GENCO                   |
| \(S_{\text{Gen}}^{t}\) | \(p_{\text{disp}}^{t}\) | The amount of dispatchable generators’ capacity                                         |
|                 | \(L_{\text{load}}^{t}\) | The amount of announced total load demand                                              |

Remark 5. Although we use the MDP to describe the upper-level market clearing procedure, in existing literatures, the upper-level is simply implemented using convex programming solvers (after being convexified) and considered as part of the environment of lower-level, which is simulated by RL algorithms. The “actions” of ISO just basically refers to the decision variables in the upper-level market clearing problem.

Table 4. Objectives and Constraints of Different Stakeholders

| Types of Stakeholders | Model      | Description                                      |
|-----------------------|------------|--------------------------------------------------|
| GENCOs                | Objective  | ✔ To maximize the total reward by participating in the market |
|                       | function   |                                                 |
|                       | Constraints| 1-a) Capacity limitation                         |
|                       |            | 1-b) Bidding price limitation                    |
1-c) Ramp capacity limitation  
1-d) Localized power balancing (for prosumers)  
1-e) Load curtailment limitation (for prosumers)  

| Objective function |  |
|---------------------|---|
| ✔ To minimize the total cost of purchasing energy in the market |  |

Constraints

| Constraints |  |
|-------------|---|
| 2-a) Network power balancing constraints |  |
| 2-b) Available generation capacity limitation |  |
| 2-c) Line capacity limitation |  |
| 2-d) Load curtailment limitation |  |

Fig. 4. Framework of Generalized Dispatching Problems

Existing literatures deploying different RL algorithms to simulate the bidding procedure and compute the Nash equilibrium point (NEP) are comprehensively summarized in Table 5, from which it can be observed that the DDPG algorithm and its modified versions are the state-of-the-art method to be adopted. Hence, we herein consider the DDPG algorithm as the baseline, to discuss its main limitations with potential solutions for further investigation.

- **Convergence:** The DDPG algorithm under a multi-agent bidding environment is always hard to be converged because of the following reasons. Recall that DDPG is implemented under the architecture of actor-critic algorithm, in which the policy function is updated in the actor network based on the “objective performance” estimated in the critic network (see Section 2.4.2). However, this estimation may not be accurate especially at the beginning stage of training, which will further render the policy improvement intractable. Meanwhile, this algorithm is sensitive to the manipulation of hyper-parameters, which is even more complicated under the actor-critic framework with several neural networks coordinated with each other. To address this problem, some more advanced neural networks are deployed to capture the hidden information in a more effective manner, for example, the Long Short Term Memory (LSTM) network [54,77,83].

- **Privacy Protection:** The DDPG algorithm under a multi-agent bidding environment always requires the information of other agents’ bidding strategies, which will not be available in the power market, because the participants’ information is confidential especially in a competitive environment. There are two main approaches to tackle this challenge: 1) The incorporation of the attention mechanism [70], which enables the privacy protection by centralized training of the critic network and distributed training of each agents’ actor network, while each actor will make their decisions (policies) without any knowledge of other agents’ confidential information. 2) The deployment of fully-distributed MADDPG algorithm [57], in which the critic network of each agent is trained using localized observation only, which will be further utilized for estimating other agents’ policy function to facilitate the critic network training.

- **Neglection of Mixed Strategy Nash Equilibrium (MSNE):** The variance and non-predictable nature of other agents’ bidding strategies implicitly add uncertainties to each agent’s reward function. Hence, the DDPG algorithm, which is intended to generate the deterministic optimal policy, cannot capture the nature of the objective to maximize the expected total reward under uncertainties. In addition, considering that the bidding game is a long-term and repetitive procedure, GENCO agents will be motivated to randomly change their bidding strategies to prevent it from being observed or estimated by others. Hence, it is more practical to consider the final converged point as the
MSNE, which differentiates from conventional NE, because the MSNE is a probabilistic distribution, rather than a deterministic function.

| Ref No. | Year | Algorithm | Market Type | Brief Description |
|---------|------|-----------|-------------|-------------------|
| [54]    | 2022 | Soft actor-critic (SAC) | Pool-based | This paper proposes a novelMulti-task Deep RL approach based on SAC algorithm for the problem of joint bidding and pricing optimization, in which the LSTM is used to distill the temporal characteristic. |
| [55]    | 2022 | Model-based asynchronous advantage actor-critic (MB-A3C) | Pool-based | This paper adopts resilient MB-A3C method to generate bidding policies for wind energy, and its outstanding performance of optimizing profits and overcoming the uncertainties in markets is demonstrated by comparison with conventional DDPG algorithm. |
| [56]    | 2022 | Q-Learning | Pool-based | This paper adopts Q-learning to solve the bi-level problem formulated by the interaction of GENCOs and electricity retailers, where the upper-level maximizes profits for individual participants and the lower-level clears the market with uniform prices by Lagrange relaxation method. |
| [57]    | 2022 | SAC, twin delayed deep deterministic policy gradient (TD3) | Pool-based | This paper employs SAC and TD3 algorithms to generate optimal strategies for VPPs and electric vehicles (EVs) charging stations respectively in a Stackelberg game model, consisting their individual interests and the interaction among them. |
| [58]    | 2022 | Q-learning | Pool-based | This paper proposes a novel distributed online RL algorithm to solve the decision-making problem for customers and GENCO simultaneously with information exchange between supply and demand sides. |
| [59]    | 2021 | DDPG | Pool-based | This paper proposes a quarter-hourly dynamic pricing strategy based on DDPG algorithm, to address the discrete problem of traditional time-sharing pricing model. |
| [60]    | 2021 | MADDPG | Pool-based | This paper creates a non-cooperative and cooperative model among the thermal GENCOs and use MADDPG to solve it, and to explore the bidding strategy for these GENCOs under medium and long term market changes. |
| [61]    | 2021 | Double Deep Q-Networks (DDQN) | Pool-based | This paper proposes a novel energy trading system among MGs, with a DDQN algorithm and a double Kelly strategy developed for improving MGs’ profits in the bidding game while reducing dependence of main grid. |
| [62]    | 2021 | Function approximation-based Q-Learning | Pool-based | This paper firstly proposed a bidding model for the battery energy storage system (BESS) in the automatic generation control (AGC) and energy market, then solved the bidding problem with the RL, in which the trick of function approximation is adopted to avoid the curse of dimensionality. |
| [63]    | 2021 | Multi-Agent Deep Q-Network (MADQN) | Pool-based | This paper proposes a market mechanism for double auctions of regional MGs, with a MADQN algorithm to find the optimal bidding strategy for these MGs. Besides, an optimal equilibrium selection (OES) is proposed to ensure the benefit fairness, execution efficiency, and privacy protection in the interactive learning process of MADQN. |
| [64]    | 2021 | Rainbow Deep Q-Networks | Pool-based | This paper adopts the Rainbow Deep Q-Networks is used to control a battery in a MG to perform energy arbitrage in the bidding game and to improve the utilization of solar and wind energy sources. |
| [65]    | 2021 | WoLF-PHC | Pool-based | This paper adopts the WoLF-PHC algorithm to obtain the NE of the bidding game among multiple MGs in an active distribution network. The motivation of each MG’s bidding strategy convergence is analyzed using the evolutionary game theory. |
| [66]    | 2021 | MADDPG | Pool-based | This paper use MADDPG algorithm to find the NE in the MDP of bidding games, where GENCOs try to develop the optimal bidding strategy in DA market with limit information. |
| [67]    | 2021 | DDPG | Pool-based | This paper utilizes DDPG algorithm to maximize the rewards in a two-stage market for active distribution network, constituting the zonal power market and the local flexibility market. |
| [68]    | 2021 | Modified Erev-Roth algorithm | Pool-based | This paper adopts RL algorithms to help electricity consumers to make trading decisions in short-term local power market, considering the participation of demand response, share of renewable generation and degree |
The possible influence on the revenue of Wind with long short delayed reward method Peer-to-Peer

This paper proposes a novel P2P transactive trading scheme based on the MAAC algorithm, which incentivizes prosumers to engage in local energy trading while also penalizes each prosumer’s addition to rebound peaks.

This paper provides a novel hybrid community P2P market framework for multi-energy systems, in which a market surrogate model-enabled DDPG method is proposed to facilitate P2P transaction within technical constraints of the community delivery networks.

This paper proposes a modified Intelligent Priority Selection based RL (IPS-RL) method to detect false data injection attack (FDIA) in P2P energy trading online, and compare it with other methods like support vector machine (SVM), RL and its variants.

This paper models the P2P energy trading in the DA market as a MARL problem, handled by a novel MADDPG algorithm abstracting the other agents’ observations and actions.

This paper tries to solve the high dimensional continuous problem of energy trading and conversion problems for multi-energy microgrids by the employment of novel MARL approach, which combines MA-AC with the twin delayed DDPG algorithms.

This paper adopts DDPG algorithm to model the bidding strategies of GENCOs. The proposed model can also reflect the different levels of competition and find the critical value that triggers tacit collusion by quantitatively adjusting the discount factor.

This paper proposes a novel DDPG method with a prioritized experience replay (PER) strategy, enabling market participants to receive accurate feedback regarding the impact of their bidding decisions on the market clearing outcome.

This paper focuses on the bidding strategy of battery participating in the energy arbitrage market. A noisy network based RL approach is applied to learn the optimized control policy for storage/discharge, while a hybrid Convolutional Neural Network (CNN) and LSTM model is adopted to predict the price for the next day.

This paper proposes a Multi-agent-based-DQN to find the optimal bidding strategy to ensure EV owners obtaining more economic benefit and spending less time on charging.

This paper investigates the possible influence on the revenue of Wind Power Producer if they take part in both the energy and reserve market, by using the asynchronous advantage actor-critic (A3C) method to simulate the bidding procedure and compute the NE.

This paper combines the deep Q-network based algorithm and long short-term delayed rewards method to devise better trading strategy in P2P trading by evaluating delayed reward.

This paper adopts the WoLF-PHC to obtain the NE of a pool-based energy market consisting of wind turbines and EV aggregators, considering the impact of uncertainty of power outputs on their bidding strategy preferences.

This paper presents a deep RL-based scheme for trading energy of autonomous MG with other MGs and the upstream power plant, in order to increase the MG utility.

This paper proposes a new RL method to enable GENCOs to get accurate reward form the environment in multiple dimensional continuous states and action spaces, regarding their policy impact on the market clearing outcome, and thereby facilitates more accurate equilibrium analysis.

| Year | Authors | Method | Type |
|------|---------|--------|------|
| 2021 | SAC     | Pool-based |
| 2021 | Multi-Actor-Attention-Critic (MAAC) | Peer-to-Peer |
| 2021 | DDPG    | Peer-to-Peer |
| 2021 | SARSA   | Peer-to-Peer |
| 2021 | MADDPG  | Peer-to-Peer |
| 2021 | MA-TD3  | Peer-to-Peer |
| 2020 | Episode-dependent DDPG | Pool-based |
| 2020 | DDPG with Prioritized Experience Replay (PER). | Pool-based |
| 2020 | NoisyNet-DDQN (NN-DDQN) | Pool-based |
| 2020 | MADQN   | Pool-based |
| 2020 | Asynchronous Advantage Actor-Critic (A3C) | Pool-based |
| 2020 | DQN with long short-term delayed reward method | Peer-to-Peer |
| 2019 | WoLF-PHC | Pool-based |
| 2019 | DQN     | Pool-based |
| 2019 | MADDPG with long short term memory (LSTM) | Pool-based |

This paper employs SAC algorithm to construct a coordinated bidding/operation model for the wind farm improving their profits, regarding the cooperation of bidding and energy storage system (ESS).
| Year | DQN Information |
|------|-----------------|
| 2019 | DQN | Peer-to-Peer |
|      | This paper modifies DQN method to help MGs devise better trading strategies in local energy market, through much efficiently utilizing their sources and batteries. |
| 2018 | DQN | Pool-based |
|      | This paper proposes prosumers’ energy trading behavior, with the operation of an energy storage system (ESS), in a novel event-driven local energy market. The optimal bidding strategy of these prosumers is obtained by deploying the DQN algorithm. |
| 2017 | Bush–Mosteller (B–M) learning | Pool-based |
|      | This paper developed an adaptive learning algorithm, which is used to generate the strategy probability distribution for seeking the MSNE, in constrained energy trading games among individually strategic players with incomplete information. |
| 2016 | Fuzzy Q-Learning | Pool-based |
|      | This paper adopts the fuzzy Q-learning approach for the hour-ahead power market bidding procedure simulation in the presence of renewable resources. |
| 2016 | Double Deep Q-Networks (DDQN) | Pool-based |
|      | This paper propose a RL based MG energy trading scheme that applies the deep Q-network (DQN) to improve the utility of the MG for the case with a large number of the connected MGs, and reduce the dependence on power plants. |
| 2012 | Deterministic Policy Gradient (DPG) | Pool-based |
|      | This paper compares the Deterministic Policy Gradient (DPG) algorithm, which uses artificial neural networks for policy function approximation, with traditional value function based methods in simulation of bidding game among GENCOs. |

### 3.1.2 Bi-level Dynamic Pricing Considering Demand – Side Management (DSM)

The demand side management, as an emerging and effective facilitator of network real-time balancing in the de-regulated power system, refers to the incorporation of demand-side users who are financially incentivized to deduct their load demand while the electricity price in the wholesale market is extremely expensive or the load demand cannot be satisfied. There are two main approaches of DSM, including the incentive-based method and price-based method. The former refers to the fixed demand-side participants who are selected by bidding competition, to provide required balancing support following the ISO’s instruction, while remunerations will be issued based on their actual provided load deductions. The latter, which is the main focus of this subsection, denotes the dynamic pricing in the retail market with a higher price at peak-load time and lower price in the valley period. Intuitively, how to exactly determine the price to incentivize the demand-side users with a minimized cost but ensuring the load deduction is the key issue to be tackled.

![Fig.5. Framework of Generalized Dynamic Pricing Problems](image-url)
Such a bi-level DP problem can be considered as a sequential Stackelberg game, in which the ISO denotes the leader who imposes the selling and purchasing price to the demand-side users, which may constitute customers (loads) and prosumers (EV Aggregators and MGs). Then, the demand-side users (as followers) determine the load curtailment and generation schedules based on the imposed incentive price. Hence, the upper level of this model refers to the ISO’s optimal pricing problem, with the objective function indicating the maximization of cumulative expected profit (discussed in Section 2.1.2) of selling/purchasing energy to/from demand-side users, while actions satisfying certain constraints (summarized in Table.7). Note that the lower level problem is embedded in the upper level’s state constraints, indicating that the state of upper level problem is partially obtained from the optimal solution of the lower level problem, which intends to maximize the cumulative expected profit via load curtailment and selling energy to the ISO. The numerical expression of such dynamic pricing model is formulated in (28), and the elements of MDP are detailed in Table.6.

**Upper Level:**

$$\text{max}_{\mu_{\text{ISO}} \in \mathcal{K}_{\text{ISO}}} \mathbb{E} \left[ \sum_{k=0}^{T-t} \gamma^t R_{\text{ISO}}^{t+k} \mid S_{\text{ISO}}^t \right]$$

s.t.

$$A_{\text{ISO}}^t \in \mathcal{A}_{\text{ISO}}$$

$$\hat{S}_{\text{ISO}}^t = \text{arg max}_{\mu_{\text{Load}} \in \mathcal{K}_{\text{Load}}} \mathbb{E} \left[ \sum_{k=0}^{T-t} \gamma^t R_{\text{Load}}^{t+k} \mid \mu_{\text{Load}} \sim S_{\text{Load}}^t \right] \in S_{\text{ISO}}$$

**Lower Level:**

$$\text{max}_{\mu_{\text{Load}} \in \mathcal{K}_{\text{Load}}} \mathbb{E} \left[ \sum_{k=0}^{T-t} \gamma^t R_{\text{Load}}^{t+k} \mid \mu_{\text{Load}} \sim S_{\text{Load}}^t \right]$$

s.t.

$$A_{\text{Load}}^t \in \mathcal{A}_{\text{Load}}$$

### Table 6. MDP Elements of Dynamic Pricing Problem

| Elements    | Details and Descriptions                  |
|-------------|-------------------------------------------|
| $A_{\text{ISO}}^t$ | Price of selling energy to demand-side loads |
| $A_{\text{ISO}}^{t,buy}$ | Price of purchasing energy to demand-side prosumers |
| $P_{\text{comp}}^t$ | The amount of controllable generators’ output |
| $S_{\text{ISO}}^t$ | The amount of renewable generators’ output |
| $P_{\text{Load}}^t$ | The amount of total load demand |
| $A_{\text{Load}}^t$ | The amount of power supply from demand-side prosumers |
| $S_{\text{Load}}^t$ | The amount of demand of controllable loads |
| $L_{\text{net}}^t$ | The amount of total net load at the demand side |

### Remark 6.
To obtain the expression of constraints for the lower level problem, additional ancillary variables and equations are required to link the elements of $A_{\text{Load}}^t$ with parameters which are already known. Considering these equations are developed by capturing specific operation characteristics of different demand-side users, which are not the main focus of this review, we herein only briefly summarize some of their annotations (listed in Table.7), while detailed formulations are available in [90] – [102].

### Table 7. Ancillary Equations for Operation Status/Constraints Modelling

| Demand-side Users    | Ancillary Equations for Operation Status/Constraints Modelling                              |
|----------------------|---------------------------------------------------------------------------------------------|
| ESS/EV Aggregator    | Amount of power charging and discharging to the main grid                                  |
|                      | Cost of charging and discharging                                                           |
|                      | Discrete status of charging and discharging                                                |
|                      | Limitations of charging/discharging rates                                                  |
|                      | Status of connection/disconnection to the main grid                                         |
Table 8 provides a comprehensive summary of existing literatures deploying RL algorithms to solve such a bi-level dynamic pricing problem, with some key challenges as discussed in the following:

- The MDP is partially-observable (denoted as POMDP) due to the randomness of demand-side users’ behaviors [92] (e.g., the arrival and departure time of EVs) and the concern of privacy protection (e.g., the users’ habitual preference and daily routine can be inferred from their behaviors of using smart-home devices), which renders the implementation of conventional RL algorithms intractable. There are several applicable approaches to tackle this problem: 1) The vectorized advantage actor-critic (VA2C) algorithm [92], in which the incorporation of recurrent neural network (RNN) effectively facilitates the time-series prediction of the total loss gradient, by using the internal memory to extract the significant but hidden features of the MDP. 2) The reformulation of such a POMDP as a product belief MDP (PBMDP) [153], which is induced from the probabilistic estimation of the states from the historical data of observations, while the memory dependency is relaxed by the product technique, and therefore the obtained POMDP can be considered as fully-observable. 3) The imitation learning and inversed reinforcement learning, which will be discussed latter in Section 4.

- There will be several potential obstacles for the convergence of designated algorithms: 1) the randomness of both renewable distributed generators (RDGs) and downstream MGs, which brings uncertainties to the reward function. 2) The “curse of dimensionality” caused by the large dimension of state-action pairs. To tackle the first problem, a comprehensive introduction of risk-averse RL algorithms as the most feasible solution is provided in Section 4. For the second problem, on the one hand, the policy-based RL algorithms are recommended due to its better performance on the scalability, as discussed in Section 2.4; on the other hand, the re-formulation of elements in the state-action pair and experience tuple is proved to be an effective trick in [91], in which both the complexity of computation and burden of memory are significantly reduced.

- Last but not the least, the ISO does not know exactly the impact of incentive pricing on the demand-side users’ scheduling preferences. One possible approach is to consider the response of demand-side users as a tuple of stochastic variables, and then deploy the risk-averse RL algorithms introduced in Section 4; however, this method may lead to sub-optimal solutions, because the objective of those RL algorithms is to mitigate the risk which are inevitable, i.e., cannot be avoided by the improvement of prediction accuracy, in a “conservative” manner with more preference on the robustness to risks instead of pursuing more profits. The most applicable method, as proposed in [98], is to synthesize the users’ behavior prediction into the framework of bi-level dynamic pricing problem, in which the prediction is implemented by adopting the LSTM network. Compared with conventional neural networks, the LSTM framework enables better utilization of historical memory by adding additional components of forget gate, input gate and output gate to mitigate the problem of gradient disappearance and gradient explosion [154]. Interested readers may consult [155] for a more contextualized understanding of LSTM and some modified versions with even much better performance.

| Table 8. Summary of Existing Literatures of Dynamic Pricing |
|---------------------------------|-------------------------------------------------|-----------------|
| **Number** | **Year** | **Algorithm** | **Brief Description** |
| [90] | 2021 | Robust Adversarial Reinforcement Learning (RARL) | This paper builds a real time demand response model considering integration of renewable energy and effective demand response analysis, while the optimal strategy for consumers’ scheduling is obtained using RARL and Gradient Based Nikaido-Isoda Function to protect the consumers' privacy. |
| [91] | 2021 | Multi-Agent Q-Learning | This paper proposes a conceptual architecture to support DR management of a diversified facility in the context of a price-based DR environment. Then, a multi-agent Q-learning algorithm is developed to compute the NE of bidding strategy. |
| [92] | 2021 | Vectorized advantage actor-critic (VA2C) | This paper adopts the VA2C algorithm with the incorporation of recurrent neural network (RNN), which effectively facilitates the time-series prediction, to address the load schedule in a residential MG with consideration of privacy protection. |
| [93] | 2020 | SARSA and Q-learning | This paper modifies the conventional SARSA and Q-learning method for efficient demand response, by representing the reinforcement learning elements in a manner that |
This paper adopts the DDQN algorithm to solve the MDP of demand response management problem, achieving the goals of both regulating the voltage and reducing the total operation costs of DSO under variable electricity consumption patterns.

This paper proposes a new energy management system for demand response by using RL and fuzzy reasoning (FR), considering human preference by directly integrating user feedback into its control logic.

This paper proposes an incentive-based demand response program, in which Q-learning method was utilized to explore the optimal incentive rates at each hour which can maximize the profits of both energy service providers (ESPs) and EUs.

This paper proposes a DR scheme for energy management of discrete manufacturing systems, which is formulated as a partially-observable MDP, and then solved by a MADDPG algorithm, to obtain the optimal schedule strategy of different customers.

This paper proposes a pricing method that combines LSTM networks and RL to solve the pricing problem of service providers when the customers’ response behavior is unknown.

This paper formulates the explicit DR bidding model with the concept of agility, and then proposed a novel Q-learning algorithm to compute the NE in a decentralized fashion.

This paper proposes a dynamic pricing of DR in a hierarchical decision-making framework as a discrete finite MDP, and Q-learning is adopted to solve this problem.

This paper models the users’ long-term load scheduling problem as an MDP, while an online load scheduling learning (LSL) algorithm based on the actor-critic method is developed to determine the users’ Markov Perfect Equilibrium (MPE) policy.

This paper proposes a novel energy management system (EMS) formulation with consideration of DR, which is solved by the Q-Learning to obtain the optimal scheduling strategies of participated customers.

In this section, two important applications of RL on power market operation and management are introduced, including the NEP estimation in the environment of multiple GENCOs bidding game and ISO – demand-side users’ interactions, in order to provide theoretical guidance to the market regulation development for ISO and bidding strategy optimization for GENCOs. Here, as a summary, a comprehensive comparison of all the RL algorithms deployed in cited literatures is made in Table.9, in terms of their types, applicability in different environments and major insights, to help readers select the most appropriate algorithm based on their own circumstances and requirements.

### Table 9. Comparison of Deployed RL Algorithms in Bidding Problems

| Algorithm | Type | Applicability | Brief Introduction & Major Insights |
|-----------|------|---------------|-------------------------------------|
| Q-learning | ✔️ | ✔️ ✔️ ✔️ ✔️ | (See Section 2.3) |
| Fuzzy Q-learning | ✔️ | ✔️ | • An extension of baseline Q-learning  
• Fuzzy rules in Q-learning to deal with fuzzy environment, i.e., fuzzy goals and constraints. |
| SARSA | ✔️ | | • On-policy  
• The behavior and target policy are all \( \epsilon \)-greedy  
• Employment of bootstrapping method |
| AC | ✔️ ✔️ | | • A combination of policy-searching-based RL and value-based RL  
• The Actor network uses policy-searching method for updating the target policy  
• The Critic network uses value-based method for updating the Q-value, to compute the loss function for updating the actor network |
| A3C | ✔️ ✔️ | | • An extension of baseline actor-critic  
• Employment of multiple asynchronous actors with different behavior policies, to mitigate the correlation of sampled data.  
• A global actor network is adopted, which is coordinately updated by multiple actors. |
Although a number of fancy RL algorithms have been proposed and deployed, state-of-the-art researches are still far away from successful applications in the real world, with the following problems to be further investigated:

- The NEP is sensitive to specific market regulations – even slight modifications of bidding rules may lead to changes of each GENCO’s reward function, which would result in huge differences in term of NEP. Existing literatures focus more on the methodology of deploying RL algorithms to compute the NEP, while the modelling of bidding game environment is somehow over-simplified with the following concerns neglected, which are significant and emerging features in future power markets:

  1) The incorporation of GENCOs with RDG: Firstly, RDGs are potential price-takers who always tend to submit lower bidding price because of their extremely-low generation cost; therefore, the market clearing price may be apparently decreased, which have been observed in many existing real-world power markets [2]. Secondly, the inaccuracy of RDG output will result in differences between DA dispatch and RT operation, while the penalty imposed for such differences may affect their bidding strategy preference.

  2) The incorporation of congestion management: The increasing energy demand and lack of transmission line capacity expansion will lead to network congestions [156] with endogenized congestion surplus, while the allocation of such congestion surplus (i.e., congestion management) may result in motivations for GENCOs to intentionally “create” the congestion by abusing the market power [157] to manipulate the market clearing price; therefore, in such circumstances, the NEP will be apparently different from those cases
considering the energy trading only.

3) The coupling among energy, ancillary service and carbon emission auction markets: The incorporation of carbon emission auction market (CEAM) \([158]\) and ancillary service market (ASM) is an emerging trading paradigm, which not only promotes the elimination of carbon emission, but also facilitates the secure operation of power network in support of the deployment of frequency and voltage regulation. Hence, while these two markets are simultaneously considered, GENCOs should make the optimal allocation of their available generation capacities, to reserve certain capacity for participating in the ASM, and to consider the benefit for selling/purchasing carbon emission quotas in the CEAM.

However, the MDP model will be extremely complicated if all the market regulation details mentioned above are taken into consideration simultaneously, which will result in overwhelming computational burdens for the implementation of RL algorithms. Hence, it is necessary to conduct the sensitivity analysis of the impact of different market regulation on the bidding strategy preference of GENCOs.

- Even when the sensitivity analysis is done and the computational burden is mitigated, the obtained NEP still cannot provide clear insights on factors that motivate GENCOs to abuse the market power or to hinder the fairness of competition, because the ISO needs to “detect” such irregular behaviors by exploring hundreds of (even thousands of) possible NEPs under different load demands and market regulations. A more efficient approach is to explicitly identify the key factors that will affect the GENCOs’ bidding strategy preferences in a more macroscopical manner, and thereafter deploy the RL method for detailed validations. In \([159]\), some initial attempts have been made to adopt the Evolutionary Game Theory (EGT) with dynamic replication equations (RDEs), which are numerically related to MARL algorithms, to identify such key factors that affect the GENCOs’ decisions while considering the congestion management. Further investigations are therefore still required to deploy this methodology under other emerging market paradigms.

3.2 Optimal Dispatching in Transmission and Distribution Networks

In this section, two main paradigmatic approaches of RL-based dispatching are introduced, including centralized dispatching in both the transmission and distribution network, and distributed (parallel) dispatching of distribution network with autonomous downstream MGs and VPPs. For each of them, we will start with a generalized MDP formulation followed by a summary of existing literatures and RL algorithms deployed. Then, possible obstacles while implementing these algorithms will be discussed with solutions extracted from recent works.

3.2.1 Centralized Economic Dispatching Using RL in Transmission and Distribution Networks

In order to provide a paradigmatic and recapitulative summary of centralized ED problem covering most of existing literatures with mathematical models and corresponding RL-based solutions, we herein consider the demonstrative market framework shown in Fig.4, in which the ISO will coordinate the power dispatching of both the upstream generators and downstream autonomous VPPs and MGs, in order to minimize the expected total cost \(C_{ISO}\) while satisfying all the necessary constraints. The MDP of such an optimal dispatching problem can therefore be formulated as \((30)\), with elements detailed in Table.10:

\[
\min_{\mu_{ISO} \in \mathcal{A}_{ISO} \mid S_{ISO}^t} E \left[ \sum_{t=0}^{T-1} \gamma^t C_{ISO}^t \mid S_{ISO}^t \right] \\
\text{s.t.} \\
A'_{ISO} \in \mathcal{A}_{ISO} \\
\text{network constraints.}
\]

(30)

Note that \((30)\) is very similar to \((28)\) with the following differences to be emphasized: firstly, we mainly focus on how the ISO will optimally allocate generation schedule for each participated generators, and therefore strategic bidding decisions of generators are neglected, while we simply assume the bidding price equals to the real generation cost; secondly, the inverted energy procurement from downstream stakeholders to the main grid will also be neglected, because the endogenized energy pricing problem will further complicate the formulation of optimal dispatching. Existing literatures focusing on this topic are summarized in Table.11 in terms of involved stakeholders and deployed RL algorithms.

| Table 10. MDP Elements of Centralized Dispatching Problem |
|-----------------|---------------------------------------------------------|
| Elements | Details and Descriptions |
| \(A_{ISO}^t\) | \(A^t_{ISO,mar}\) The computed market clearing price |
| | \(p_{Gen,allo,i}^t\) The amount of allocated generation to the \(i\)th controllable generator |
| | \(\tau_{RDG,j}^t\) The on-and-off status of the \(j\)th renewable generator |
| | \(\tau_{switch,g}^t\) The on-and-off status of the \(g\)th switch in the power network |
Adding, using the solution of DDQN and multiple MGs as a Stackelberg game model, while the MADDPG algorithm is - ance the - - - y

| Constraints | a) Network power balancing constraints b) Available generation capacity limitation c) Line capacity limitation d) Load curtailment limitation e) Other operational constraints of generators, ESSs and EVs |

| Table 11. Summary of Existing Literatures of Centralized Dispatched Problem |
| Number | Year | Algorithm | Brief Description |
|---------|------|-----------|-------------------|
| [103]   | 2022 | DQN       | This paper proposes a DA optimal dispatching strategy for active distribution network based on modified deep reinforcement learning with several tricks, in order to address the challenge of uncertainties in the dispatching problem. |
| [104]   | 2022 | PPO       | This paper proposes a RT optimal energy management strategy for MG, in which the PPO algorithm is adopted to learn from historical data in the training procedure, to capture the uncertainty characteristic of renewable energy generation and load consumption. |
| [105]   | 2022 | DDQN+DDPG | This paper proposes a distribution network reconfiguration strategy with the objective of minimizing both the operation cost and load shedding, using the solution of DDQN and DDPG separately in its three-stage optimization manner. |
| [106]   | 2022 | SAC+TD3   | This paper develops a Stackelberg game model between the VPP and downstream EV cluster, and deploys the SAC and TD3 algorithm to obtain their scheduling strategies. |
| [107]   | 2022 | Swarm intelligence (SI) based DDPG | This paper considers the generation control and the optimal dispatch in an integrated manner, with the solution of SI-based DDPG algorithm to improve the robustness in the real-world implementations. |
| [108]   | 2022 | DDPG      | This paper proposes an optimal dispatching strategy for the integrated energy systems, by deploying the DDPG algorithm with several tricks to improve its training performance. |
| [109]   | 2021 | Q-learning | This paper effectively integrates the restoration and optimal dispatching to enhance the power grid resilience, while the Q-learning is adopted to generate the sequential order of repairing damaged components and update the network topology. |
| [110]   | 2021 | Q-learning | This paper develops a RT dispatching strategy for the distribution network using the Q-learning algorithm, with the objective to avoid unexpected failures and decrease the operation and maintenance costs. |
| [111]   | 2021 | DQN       | This paper proposes an intra-day dispatching strategy considering the uncertainty of net load, using the DQN algorithm to improve the overall performance. |
| [112]   | 2021 | TD3+experience pool replay (REPR-TD3) | This paper proposes an automatic generation control (AGC) dispatch strategy using the modified TD3 algorithm with experience pool replay, in order to reduce the area control deviation and regulation mileage payment. |
| [113]   | 2021 | MADDPG    | This paper develops a two-stage Volt-var optimization method using MADDPG to dispatch the on-load tap changer (OLTC), capacitor banks (CBs) and reactive power generations, to mitigate fast voltage violation while minimizing the network power loss. |
| [114]   | 2021 | HMO-DDPG  | This paper proposes a hierarchical multi-objective deep deterministic policy gradient (HMO-DDPG) algorithm for economic dispatching, considering the security, economy and environmental protection simultaneously. |
| [115]   | 2021 | MADDPG    | This paper proposes a bi-level energy management framework for the distribution system with multiple MGs as a Stackelberg game model, while the MADDPG algorithm is deployed to obtained the optimal dispatching strategy of the main network and MGs. |
| [116]   | 2021 | Finite-horizon DDPG (FH-DDPG), finite-horizon recurrent DDPG (FH-RDPG) | This paper proposes two novel algorithms, namely the finite-horizon DDPG (FH-DDPG), and finite-horizon recurrent DDPG (FH-RDPG), to obtain the optimal dispatching strategy of MG with and without fully observable state information, which refers to the impact of net load uncertainty on the observation of state transitions. |
| [117]   | 2021 | Differential Evolution (DE) + Q-learning | This paper combines the conventional the Differential Evolution (DE) with Q-learning to solve the economic dispatching problem, in which the Q-learning is deployed to facilitate the task of parameter control for DE. |
| [118]   | 2021 | DDPG      | This paper proposes a scenario-based robust economic dispatch strategy for virtual power plants, using the generative adversarial network (GAN) for data augmentation and DDPG. |
This paper develops an energy management framework of an MG consisting of several flexible resources, and seven RL algorithms are deployed to examine their performances.

This paper proposes a multi-objective energy management framework of a hybrid energy system considering the environmental friendliness, deploying PPO as the solution.

This paper provides two applicable methods to tackle the non-observability in the energy management procedure, which is due to the uncertainty of net-load and electricity price, by incorporating recurrent neural networks and variational inference into the framework of AC algorithm.

This paper considers the tie-line power schedule and the economic dispatch in a regional power system with uncertainty of net load, while a novel simulated annealing Q-learning algorithm is proposed to avoid the algorithm trapping into local optimality.

This paper presents an intelligent power flow adjustment and dispatching model based on SARSA algorithm to replace the manual intervention.

This paper proposes a method to solve real-time optimal dispatching through a Lagrangian based on DDPG algorithm, while incorporating the operational constraints into the DDPG using the intuition of Lagrangian relaxation.

This paper proposes an accelerated multi-objective RL method for the coordinated dispatching of generators and compensators in the power system, with the objectives of minimizing the expense of control decision and the voltage deviation.

This paper proposes an online network reconfiguration scheme based on the DQN approach, in which the distribution network operator modifies the network topology to change the power flow when the reliability of network is threatened.

This paper proposes a Q-learning algorithm augmented with Monte-Carlo tree search (MCTS) to address the problem of dispatching of a battery energy storage system (BESS) in the MG.

This paper focuses on the dynamic optimization problem of automatic generation control (AGC) with considerations of regulating performance, generation cost and emission, based on multi objective Q-learning (MOQL) and small population-based particle swarm optimization (SPPSO).

This paper deploys the Q-learning algorithm to obtain the strategy of ancillary service (peak regulation) procurement provided by VPPs.

This paper develops a multi-objective optimization framework of the energy internet, and the objectives include minimizing power generation cost, lifetime extension for BES, and realizing a rational operation principle, with the solution of A3C algorithm.

This paper proposes a novel RT energy management method of an MG considering the uncertainty of the load demand, renewable energy, and electricity price, while the DQN algorithm is deployed to obtain the optimal dispatching strategy.

This paper presents a model-free-based RL method for the energy management of MGs in the distribution network, while the objective of the DSO is to decrease the demand-side peak-to-average ratio (PAR), and to maximize the revenue of selling energy to MGs.

This paper adopts the DDQN algorithm with some tricks to solve the problem of active power optimal dispatching.

This paper proposes a novel bacteria foraging Q-learning with knowledge transfer method for risk-based economic dispatch, which is integrated with risk assessment theory to indicate the uncertainties of net load and contingencies.

This paper proposes a Q-learning based optimal dispatching strategy of a battery in the MG, in which the battery’s charge and discharge efficiencies, and the nonlinearity in the microgrid due to the inverter’s efficiency are taken into account.

This paper proposes a Q-learning based optimal dispatching strategy of an MG, in which the reward function is implemented by fuzzy system to improve the learning efficiency.

This paper proposes a Q-learning based optimal dispatching strategy of a battery in the MG, with the objectives of maximization of battery utilization in the peak load time, and to increase the utilization of wind turbine in the MG.
3.2.2 Decentralized Economic Dispatching Using RL in Distribution Networks with Downstream Stakeholders

Recent trends in the de-regulation of modern power systems, especially the popularization of active distribution networks (ADNs) with downstream autonomous stakeholders (DAS) including VPPs and MGs, have led to a proliferation of studies that concern the operation of such paradigm in a distributed manner, in which the downstream stakeholders independently conduct the self-dispatch while the DSO acts as a coordinator to maintain the network real-time balancing without intervention, i.e., the re-dispatch to facilitate the network balancing is always “very slight” and not compulsory. Compared with conventional centralized dispatching, such a method has advancements in the efficiency of data transmission and privacy protection, which are main reasons for the necessity of further investigations. However, in order to deploy RL algorithms to obtain optimal dispatching strategies of each DAS, the distributed computing techniques are necessary to provide the tool for the communication and coordination among DASs and the DSO. In the following, we will provide a brief review of some selective emerging dispatch paradigms using RL in a decentralized manner, facilitated by several state-of-the-art distributed computing techniques including: 1) edge computing (EC); 2) federated learning (FL); and 3) hierarchical reinforcement learning (HRL) with knowledge-guiding and data-driven model (KDM). These techniques are selected to be discussed because they are representative approaches to demonstrate how distributed computing architectures are incorporated to the implementation of RL algorithms to facilitate the dispatching of ADN with dispersed DASs. Note that the focus of subsequent elaborations will be the foundations of these mentioned techniques, as well as the framework of dispatching paradigms with incorporation of distributed computing architectures, while details of corresponding RL algorithm implementations are available in [138] – [144]. Also, a brief summary of existing literatures focusing on this topic can be found in Table 12, in which the literatures are classified based on their methodologies of dispatching, according to the aforementioned three main types.

Table 12. Summary of Existing Literatures of Decentralized Dispatching Problem

| Number  | Year | Type  | Brief Description                                                                                                                                 |
|---------|------|-------|--------------------------------------------------------------------------------------------------------------------------------------------------|
| [138]   | 2022 | HRL   | This paper develops a two-level RL framework to deal with the optimal pricing and dispatching of multiple MGs, while the uncertainties of net load are addressed with several effective tricks. |
| [139]   | 2021 | FL    | This paper proposes a Monte Carlo tree search based RL method to mitigate the impact of uncertainty of net load on the optimality of energy transactions among multiple MGs, and an actor-critic-based RL method is adopted to implement the distributed dispatching. |
| [140]   | 2021 | EC    | This paper proposes an MADQN framework for the distributed dispatching of multiple MGs, considering the uncertainty of net load and potential network attacks, in which the DQN networks of MGs are centrally trained, but the dispatching is implemented in a dispersed manner. |
| [141]   | 2021 | EC    | This paper proposes a distributed form of Hysteretic Q-learning to implement the decentralized dispatching after the virtual auction, in which actual actions are obtained via the constraint projection method considering the auction results and network operation constraints. |
| [142]   | 2021 | EC    | This paper proposes an MADQN framework for the distributed dispatching of multiple MGs, in which DQN networks of MGs are centrally trained, but dispatching is implemented in a dispersed manner. |
| [143]   | 2020 | HRL   | This paper proposes a two-level RL framework for the distributed dispatching of multiple MGs, in which the upper-level agent allocates the generation capacity to distributed MGs, who will then conduct the self-dispatching using Q-learning algorithm. |
| [144]   | 2018 | HRL   | This paper develops a multi-agent Q-learning algorithm to obtain the NE among several autonomous energy resources’ dispatching strategy in a MG. |

1) RL-based Distributed Dispatching with Edge Computing (EC) Architecture

The concept of EC refers to the transformation of computing services from the central server (for example, the “cloud”) to “edges” which are adjacent to dispersed end-users with the support of smart IoT devices. Compared with conventional centralized “cloud computing”, EC can effectively mitigate the inefficient data transmission from end-side servers to the cloud, which may result in the lack of data transmission capacity and the concerning of privacy disclosure; meanwhile, the computation burden of the cloud server is offloaded to edge servers, and the large-scale parallel processing will significantly accelerate the computing. Specifically, for the problem of power dispatching in a distributed manner, the downstream VPPs and MGs do not need to provide privacy information (for example, the generation cost) to the ISO, while the ISO can implement the dispatching based on the edge-side servers with de-composed data to be processed, avoiding the requirement of algorithm scalability discussed in Section 3.1.2.
Fig. 6. Framework of Edge-computing-based RL for Decentralized Dispatching

Fig. 6 illustrates the generalized framework of distributed dispatching paradigm based on RL with edge-computing methodology incorporated. This paradigm follows the “centralized training, distributed execution” manner: in the offline-learning phase, the actor-critic networks of each DG agent are trained simultaneously by the cloud server (managed by the DSO) based on the generation and load data transmitted from edges to the DSO, while the dispatching decision of each agent are made individually in the online execution phase, in which the DSO only serves as the coordinator and no additional data transmission is required. However, it is important to note that, the autonomy (or independence) of downstream stakeholders cannot be guaranteed under this paradigm because of the centralized network training.

Fig. 7. Framework of Federated-learning-based RL for Decentralized Dispatching
2) **RL-based Distributed Dispatching with Federated Learning (FL) Architecture**

FL is an emerging distributed computing architecture to carry out machine learning among multi-participants or multi-computing nodes under the premise of guaranteeing the information security of big data transmission, protecting the privacy of terminal data and personal data, and ensuring the legal compliance. This technique is feasible to be adopted in the condition that DASs are collaboratively training the “global model” managed by the DSO for distributed dispatching, while DASs have their own private data of operational characteristics which cannot be shared with other DASs and the DSO. Particularly, the actor-critic RL algorithm can be perfectly incorporated into the FL framework to facilitate the training, in which each DAS updates its own actor network with the data exchange to the common critic network, i.e., the “global model”. The training procedure sequentially constitutes the following steps depicted in Fig.7. For each iteration, DASs will firstly conduct the localized training of their own dispatching policy with some initialized parameters of the global model. Then, DASs will send encrypted parameters of their trained networks (the actor networks) to the global server without disclosure of any additional private data, to facilitate the subsequent update of the global model, which deploys the FedSGD algorithm [160] or the adaptive moment estimation (Adam) optimization algorithm [161]. Finally, the DSO will distribute the obtained parameters of global model to DASs for their localized training in the next iteration. Compared with the aforementioned EC architecture, FL can guarantee the independence of each DAS, with little involvement in their localized training by limited and encrypted data transmission.

3) **Hierarchical Reinforcement Learning-based Distributed Dispatching with Knowledge-guiding and Data-driven Model (KDM)**

Conventional machine learning (ML) techniques, especially DL and RL, requires a large amount of sampled data indicating the priori knowledge in order to mitigate the generalization, while the data applicable for pre-training may not be sufficient in terms of dispatching. Meanwhile, the dispatching problem has the so-called “baseline requirement” for the generalization in the partial learning space (PLS), i.e., the constraints for security operation should never be violated (to fulfill the baseline requirement), while the objective of minimizing the total cost is considered with less significance (to reduce the generalization). Such problems are denoted as “Baseline Ensuring and Global Optimality with Limited Data (BeGold)”, which cannot be tackled by conventional ML and therefore requires further investigations.

The Hierarchical Reinforcement Learning (HRL) incorporated with KDM [143] provides an effective countermeasure to this problem, with the core insight of embedding the operational constraints as priori knowledge while implementing the RL algorithm; meanwhile, the hierarchical architecture of training paradigm significantly reduces the dimensionality of action space, and therefore the training performance can be guaranteed even with limited sampled data. Specifically, operational constraints are expressed in the following way:

\[
\|f(x, \theta), k_i(x, \beta_{ki})\|_{\varepsilon} \leq \kappa_i \quad (31)
\]

\[
\Rightarrow \text{if} \|f(x, \theta), k_i(x, \beta_{ki})\|_{\varepsilon} > \kappa_i, \text{then} \|f(x, \theta), k_i(x, \beta_{ki})\|_{\varepsilon} = \kappa_i \quad (32)
\]

where \(x\) and \(\theta\) denote the state and the associated parameter indicating the policy respectively; \(f(x, \theta)\) denotes the implicit constraints for the parameter set; and \(k_i(x, \beta_{ki})\) denotes the \(i^{th}\) constraint for the state set, in which \(\beta_{ki}\) represents the confidence level of such constraints when \(x\) hedges uncertainty, and \(\kappa_i\) denotes the set of threshold value. By reformulating (31) as a logic expression as (32), the priori knowledge can be transferred to a penalty term \(\mu_i(\|f(x, \theta), k_i(x, \beta_{ki})\|_{\varepsilon} - \kappa_i)\) with Lagrangian multiplier \(\mu_i\), which will be added in the formula of Q-values in order to disregard those actions that will lead to violation of constraints.

![Fig.8. Framework of HRL-based RL for Decentralized Dispatching](image-url)
Such KDM can be incorporated into the HRL framework to facilitate the dispatching for multiple DASs in a distributed manner, as depicted in Fig.8. Compared with EC and FL, in the HRL-DRM framework, the DSO as centralized coordinator will not intervene the dispatching of autonomous MGs in any implicit (the common critic network in FL will implicitly affect the policy update of each MG) or explicit (the centralized parameter training in EC) manner. Instead, the DSO will only determine the total amount of each MG’s power output using Q-learning or other value-based RL methods in the upper-level of dispatching, whilst ensuring the tie-line capacity not violating the constraint, using the embedded KDM. Then, based on the allocated amount of generation, the autonomous MGs will implement the individual dispatching using value-based RL methods, while the operational constraints are also incorporated in the form of KDM.

3.2.3 Discussion and Summary

In this section, the application of RL techniques on power system dispatching is introduced in two separate streams, including the centralized and distributed dispatching manners. Similar to Section 3.1.3, another summary of all the RL algorithms deployed in cited literatures of this section is available in Table.13.

**Table 13. Comparison of Deployed RL Algorithms in Dispatching Problems**

| Algorithm                | Type       | Applicability          | Brief Introduction & Major Insights                                                                                                                                                                                                 |
|--------------------------|------------|------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Q-learning               | Policy-    | ✔ Continuous space     | • An extension of baseline Q-learning                                                                                                                                                                                                     |
|                          | based/Value-based | Distributed training |                                                                                           | • Fuzzy rules in Q-learning to deal with fuzzy environment, i.e., fuzzy goals and constraints.                                                                                                                                 |
| Fuzzy Q-learning         | ✔          |                        |                                                                                           |                                                                                                                                                                                                         |
| Multi objective Q-learning (MOQL) | ✔        |                        | • An extension of baseline Q-learning                                                                                                                                                                                                     |
|                          | ✔          |                        | • Different objectives are indicated by multiple reward functions with weight coefficients, and the final reward is the combination of these weighted reward functions.                                                                 |
| Bacteria foraging Q-learning | ✔ ✔ ✔      |                        | • An extension of baseline Q-learning                                                                                                                                                                                                     |
|                          | ✔          |                        | • The bacteria foraging optimization technique is incorporated into the framework of Q-learning, to extend the conventional Q-learning to a multi-agent environment.                                                                             |
| SARSA                    | ✔          |                        | • On-policy                                                                                                                                                                                                                               |
|                          | ✔          |                        | • The behavior and target policy are all ε-greedy                                                                                                                                                                                            |
|                          | ✔          |                        | • Employment of bootstrapping method                                                                                                                                                                                                          |
| AC                       | ✔          | ✔                      | • A combination of policy-searching-based RL and value-based RL.                                                                                                                                                                           |
|                          | ✔          | ✔                      | • The Actor network uses policy-searching method for updating the target policy                                                                                                                                                                 |
|                          | ✔          | ✔                      | • The Critic network uses value-based method for updating the Q-value, to compute the loss function for updating the actor network.                                                                                                                                 |
| A3C                      | ✔          | ✔                      | • An extension of baseline actor-critic                                                                                                                                                                                                      |
|                          | ✔          | ✔                      | • Employment of multiple asynchronous actors with different behavior policies, to mitigate the correlation of sampled data.                                                                                                                                 |
|                          | ✔          | ✔                      | • A global actor network is adopted, which is coordinately updated by multiple actors.                                                                                                                                                              |
| SAC                      | ✔          | ✔                      | • An extension of baseline actor-critic                                                                                                                                                                                                      |
|                          | ✔          | ✔                      | • The maximum entropy is incorporated to the objective of RL, to facilitate the exploration in the training and the robustness of obtained policy.                                                                                                                                 |
| MAAC                     | ✔          | ✔                      | • An extension of baseline actor-critic to the multi-agent environment.                                                                                                                                                                          |
| TRPO                     | ✔          | ✔                      | • (See Section 2.4.1)                                                                                                                                                                                                                           |
| PPO                      | ✔          | ✔                      | • A modification of baseline TRPO method with the first-order approximation while updating the parameterized policy.                                                                                                                                 |
| DQN                      | ✔          |                        | • (See Section 2.3.3)                                                                                                                                                                                                                           |
Apart from those obstacles elaborated in Section 3.2.1 and Section 3.2.2 while implementing RL algorithms, we would like to emphasize the following concerns specifically in the context of RL being deployed to solve the optimal dispatching problem, as well as some recently developed methods as applicable countermeasures.

- **Stringent Requirements for Reliability and Interpretability of RL**: The complexity of real-time dispatching is mainly resulted from potential security concerns (for example, the switching transients caused by commutation failure in HVDC-embedded power networks), which are related to the specific network topology and operation status, can only be mitigated using preventive control executed by intervene of human-intelligence, and cannot be precisely and numerically described as bounds for the action space in the MDP. In addition, it is hard to develop the pre-training datasets to facilitate the dispatching under emergency contingencies or other faults. Unlike the environment of toy experiments or playing computer games, the sub-optimality of dispatching strategy and resulted security issues in the real-world power system will lead to massive financial losses of millions of dollars, which cannot be tolerated. Hence, the RL technique will not be actually applicable in real-world implementations until its reliability can be strictly proved, while the “reliability” can be guaranteed by the interpretability of algorithms and pre-training datasets that cover as many scenarios as possible. As observed in Table.10, most of state-of-the-art works adopt the DDPG algorithm which contains deep neural networks with characteristics of non-interpretability. To tackle this issue, large number of current works in the area of computer science has made significant contributions, including the feature map visualization of CNN [162], the interpretation of long-distance dependency using the cell state in the LSTM [155], etc.

- **Training Efficiency Caused by Uncertainties and Large Dimensionality**: As shown in Table.9, the action of ISO may include generation output and status of switches, which means the dimensionality of action space may exceed $10^5$ in actual dispatching tasks for a regional power network; therefore, the efficiency of exploration will render the convergence of RL algorithms intractable. Meanwhile, the variance of value functions due to the uncertainty of reward function caused by the non-predictability of net load will further adversely contribute to the convergence. To tackle these two issues, the work proposed by Baidu© PARL Team in the competition of “Learning to Run a Power Network Challenge” held by NeurIPS 2020 [163] can be considered as a contribution with great enlightenment, which combines the imitation learning and RL inheriting advantages of both methods, to pursue the long-term reward while learning in a very fast and effective manner using the expertise knowledge.

| Method            | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | (See Section 2.4.2) |
|-------------------|---|---|---|---|---|---|---|---|---|---|---|---|---------------------|
| DDQN              | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |                     |
| DDPG              | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |                     |
| SI-based DDPG     | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |                     |
| FH-DDPG           | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |                     |
| Lagrangian-based DDPG | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |                     |
| HMO-DDPG          | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |                     |
| REPR-TD3          | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |                     |
| MADDPG            | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | (See Section 2.5)  |

An extension of baseline DQN
- This algorithm employs double Q-networks to approximate two value functions, so as to eliminate the overestimation of Q-values.

- An extension of baseline DDPG
- The swarm intelligence optimization technique is incorporated into the framework of DDPG, to extend the DDPG to a multi-agent environment with distributed execution manner.

- An extension of baseline DDPG
- The methodology of Lagrangian relaxation is incorporated to deal with involved constraints.

- An extension of baseline DDPG
- The hierarchical MARL framework is introduced to consider multiple objectives simultaneously.

- An extension of baseline DDPG
- Double critic networks are adopted to tackle the over-estimation of Q-values.
- The actor network is updated in a delayed frequency to remove the correlation of sampled data.

- An extension of baseline DDPG
- The instability and finite-horizon problems in DDPG are solved by techniques including backward induction and time dependent actors.
4. Future Trends of RL Applications in Deregulated power markets

4.1 Risk – Averse Reinforcement Learning

As aforementioned in Sections 3.1 and 3.2, both the bidding decision of GENCOs and the dispatching decision of ISO will be affected by the uncertain environment, including the other GENCOs' bidding decisions and the net load. Decision-makers will encounter large amount of financial losses if such uncertainties are not taken into consideration. For example, GENCOs may lose the bid if they insist making aggressive bidding strategies with relatively high bidding price, and the neglect of net load uncertainty will result in the difference of energy between DA scheduling and RT procurement, i.e., the RT imbalance, which can only be addressed by deploying the balancing and ancillary service with more expensive costs. Hence, there is a pressing need to consider the risk mitigation in both ISO’s and GENCOs’ decision making. Although considerable literatures have investigated the risk mitigation method from the aspect of conventional optimization, this methodology has not been adopted to the dynamic programming problems with RL solutions.

The Risk – Averse RL (RARL), which incorporates the core methodology of robust and stochastic programming into the RL algorithms, is intended to fill this research gap; for simplicity, we hereinafter denote these two streams of RARL as “robust RARL” and “stochastic RARL”. The main difference between RARL and conventional RL algorithms lies in the formulation of value functions (or the Q-value). Specifically, for the robust RARL, the core idea is to find the policy that has the best performance under the worst scenario with the lowest reward, and the value-function can be therefore intuitively formulated as a max-min problem:

\[ V_{\mu^k,\lambda}^{k} (s_t^{k,h}) := \max_{\mu^k} \min_{R_{t+1}^{k}} \sum_{a_t \in A_{t}^{k}} \beta^k \left( a_t^{k,h} | \theta^k \right) \left( \tilde{R}_{t}^{k} (s_t^{k,h}, a_t^{k,h}) + \gamma V_{\mu^k,\lambda}^{k+1} (s_t^{k,h+1}) \right) \]  

(33)

where \( V_{\mu^k,\lambda}^{k} \) denotes the state-value function of the kth agent’s robust policy \( \mu^k \) parameterized by \( \theta^k \). \( \tilde{R}_{t}^{k} \) denotes the reward function with uncertainty, and \( \gamma \) is the discount factor. Based on (33), one can further derive the policy gradient function for the implementation of policy-searching-based RL (please consult [147] for details). For the stochastic RL, the core idea is to incorporate the conditional-value-at-risk (CVaR) into the Q-value as a penalty term:

\[ Q_{t}^{k} = \tilde{R}_{t}^{k} \left( s_t^{k,h}, a_t^{k,h} \right) + \gamma Q_{t+1}^{k} \left( s_t^{k,h+1}, a_t^{k,h+1} \right) + \lambda_{t}^{k} \left( C_{CVaR,k}^{t} - C_{risk,max} \right) \]  

(34)

Such formulation is partially inspired by the Lagrangian Relaxation [145], in which \( C_{risk,max} \) is the threshold value of potential risk, indicating the maximum risk that can be tolerated by the decision maker. Here, such constraint is re-formulated as the penalty term with Lagrange multiplier \( \lambda_{t}^{k} \), which can be updated in the following way:

\[ \lambda_{t}^{k} \leftarrow \max \left\{ 0, \lambda_{t-1}^{k} + \sigma \left( C_{CVaR,k}^{t} - C_{risk,max} \right) \right\} \]  

(35)

where \( \sigma \) can be considered as a hyper-parameter similar to the learning rate. Meanwhile, the value of \( C_{CVaR,k}^{t} \) is also updated in each iteration, using the Robbins-Monro and Kiefer-Wolfowitz algorithm, while interested readers may consult [164] for detailed information. Note that the uncertainty set with probabilistic of each possible scenario must be provided in order to estimate the value of \( C_{CVaR,k}^{t} \) in these stochastic RARL algorithms.

Due to the similarity of core methodology, the robust and stochastic RARL inherit the advantages and drawbacks of robust and stochastic programming respectively, and therefore they are applicable to solve different problems in deregulated power markets. For robust RARL, the risk can be effectively mitigated while the expected total reward may be negatively affected because of the overly conservative policy. Hence, such method is feasible to be applied to the problems with high risk sensitivity or financial losses which cannot be tolerated; for instance, power dispatching problems especially in the transmission network. Instead, for stochastic RARL, the total expected reward will be apparently higher than that of robust RARL, but it may encounter large losses in individual cases. Hence, such method is feasible to be applied to the problems with low risk sensitivity; for example, the bidding strategy or power dispatching in the distribution network.

Existing literatures focusing on developing such RARL methods are briefly summarized in Table.14, while interested readers may take this as a reference inspiring its potential applications on operations of deregulated power markets.

| Table 14. Summary of Applicable RARL Algorithms |
|-------------------------------------------------|
| Number | Year | Algorithm | Brief Description |
|--------|------|-----------|-------------------|
| [145]  | 2021 | Conditional value-at-risk Adversarial Reinforcement Learning (CARL) | This paper proposes a novel CARL algorithm, in which a Stackelburg game formulation is developed between a policy player and an adversary that perturbs the policy player’s state transition, while the player considers the CVaR in its reward function. The existence of NE in such games is finally proved |
| [146]  | 2021 | Risk-Averse Multi-Agent | This paper proposes a Risk-Averse Multi-agent Q-learning (RAM-Q) algorithm, in which the agents are seeking for achieving the robust NE, i.e., to pursue the best |
| Year | Paper | Model/Algorithm | Description |
|------|-------|----------------|-------------|
| 2021 | [147] | Conservative Offline Distributional Actor Critic (CODAC) | This paper proposes a Conservative Offline Distributional Actor Critic (CODAC) algorithm, adapting distributional RL to the offline setting by penalizing the predicted quantiles of the return for out-of-distribution actions. |
| 2021 | [148] | Risk-Averse Bayes-Adaptive Reinforcement Learning (RABARL) | This paper solves the problem of optimizing the CVaR of total return in Bayes-adaptive MDPs, by ensuring the policy being risk-averse to both the epistemic uncertainty due to the prior distribution over MDPs, and the aleatoric uncertainty due to the inherent uncertainty of MDPs. |
| 2020 | [149] | Risk Averse Distributional Policy gradient (RASDPG) | This paper incorporates CVaR in sample based distributional policy gradients (SDPG) for learning risk-averse policies to achieve the robustness against the uncertainties in the environment. |
| 2020 | [150] | Distributional Soft Actor Critic (DSAC) | This paper proposes a novel RL algorithm that combines the SAC as maximum entropy RL with distributional RL, while providing a comprehensive risk measurement metrics to enable the simultaneous consideration of different risks in decision-making. |
| 2020 | [151] | Risk Averse Value Expansion (RAVE) | This paper develops a new method for sampling in the uncertain environment while implementing the value function estimation, based on conventional model-based value expansion (MVE), and ensuring the actions being risk-averse by seeking the lower confidence bound of the estimation. |
| 2019 | [152] | Risk Averse Robust Adversarial Reinforcement Learning (RARARL) | This paper proposes an improvement of baseline Risk Averse Adversarial Reinforcement Learning (RAARL), enabling not only the protagonist seeking for risk-averse policies, but also the adversary seeking for every possible risk. By doing so, some catastrophic events with low probability and always neglected in previous methods are taken into consideration. |

### 4.2 Inversed Reinforcement Learning

Although the aforementioned RARL method can potentially mitigate the financial losses caused by uncertainties, its applicability to real-world power systems is still not proved, because all the RL algorithms are developed based on the methodology of “error-and-trial”, while sometimes such mistakes cannot be tolerated in power system operation. For example, in the power dispatching problem, the mis-operation of switches may lead to unexpected network topology and load flow change, which will result in catastrophic frequency collapse. In addition, it is overwhelmingly complicated to incorporate every detail of power system operation into the RL algorithm design, while any negligence of details may result in huge difference between “simulation” and “real-world implementation”.

The methodology of IRL [165] is therefore proposed to tackle this problem with: 1) implicit reward functions (i.e., how the agent is interacting with the environment to obtain rewards); 2) zero-tolerance of mistakes; and 3) the dataset of past-experience, i.e., how this problem is tackled by human-intelligence (the so-called “expert dataset”). The aim of IRL is therefore to estimate the explicit reward function \( R(s_t) = \omega \varphi(s_t) \) parameterized by \( \omega \), which will be used to implement the RL procedure to find a policy function with equivalent “performance” of the expert dataset. The definition of expert dataset’s performance is inspired by the value function of a given policy \( \mu(s_t, a_t) \):

\[
V^\mu(s_t) = E\left[ \sum_{t=0}^{\infty} \gamma^t R(s_t) \mid \mu \right] = E\left[ \sum_{t=0}^{\infty} \gamma^t \omega \varphi(s_t) \mid \mu \right]
\]  

(36)

We define the “performance” of a given policy as objective expectation \( J(\mu) = E[\sum_{t=0}^{\infty} \gamma^t \varphi(s_t) \mid \mu] \). Intuitively, the objective expectation of \( m \) trajectories of expert dataset can be formulated as:

\[
J(\mu_e) = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{\infty} \gamma^t \varphi(s^i_t, a^i_t)
\]

(37)

where \( \{s^i_t, a^i_t\}, i = 1, \ldots, m \) is the expert dataset. Hence, the aim of IRL can be re-expressed as to learn a policy that minimize the difference between the objective expectation of deduced policy function and the given expert dataset, until satisfying:

\[
\left\| J(\mu) - J(\mu_e) \right\|_2 \leq \varepsilon
\]

(38)

where \( \varepsilon \) is the threshold value as a hyper-parameter.

By adopting this methodology, one can obtain an explicit policy function similar to the implicit expert policy, without requirement of detailed modelling of environment and reward function, while the impact of inaccurate model on the policy function can also be effectively mitigated. The major obstacle of the applicability to real-world power systems is the
availability of expert dataset, especially in the aspect of market operation with large quantity of confidential data. However, in terms of power system dispatching, the acquiring of the expert data is relatively easier, while the dispatching problem is under a more pressing need for algorithms with error-free capability. Hence, the application of IRL method in the dispatching problem will be a more emerging topic to be further investigated.

5. Conclusions

With the reformation of power market regimes to tackle the challenges in modern power systems with high penetration of renewable energy, the optimal bidding strategy and dispatching methodology in such new paradigms are under a pressing need to be investigated, concerned by both market participants and power system operators. In this paper, the generalized methodology of deploying RL to obtain the optimal bidding and dispatching strategy is introduced. RL algorithms have demonstrated advantageous performance compared with conventional optimization methods, and its deployment in real-world application is promising but several significant problems still need to be further investigated. More advanced algorithms, including several state-of-the-art RL techniques developed in recent years, have great potential to be modified and adopted to tackle these problems.

References

[1] M. T. Ross, “The future of the electricity industry: Implications of trends and taxes,” Energy economics, vol. 73, pp. 393–409, 2018, doi: 10.1016/j.eneco.2018.03.022.
[2] J. Morales Pedraza, Electrical Energy Generation in Europe, 2015th ed. Cham: Springer International Publishing AG, 2014.
[3] C. M. Worlen, “Economic and technological aspects of the market introduction of renewable power technologies,” Boston University, 2003.
[4] E. Panos and M. Densing, “The future developments of the electricity prices in view of the implementation of the Paris Agreements: Will the current trends prevail, or a reversal is ahead?,” Energy economics, vol. 84, p. 104476, 2019, doi: 10.1016/j.eneco.2019.104476.
[5] K. Zhang, S. Troitzsch, S. Hanif, and T. Hamacher, “Coordinated Market Design for Peer-to-Peer Energy Trade and Ancillary Services in Distribution Grids,” IEEE transactions on smart grid, vol. 11, no. 4, pp. 2929–2941, 2020, doi: 10.1109/TSG.2020.2966216.
[6] G. Zhang, J. Lu, and Y. Gao, “Bi-level Programming for Competitive Strategic Bidding Optimization in Power markets,” in Multi-Level Decision Making, Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 315–324.
[7] D. Palit and N. Chakraborty, “Constrained Optimal Bidding Strategy in Deregulated Power market,” in Artificial Intelligence and Evolutionary Algorithms in Engineering Systems, vol. 325, New Delhi: Springer India, 2014, pp. 863–873.
[8] K. Zhang, X. Wang, and S. Zhang, “Equilibrium Analysis of Power market with Wind Power Bidding and Demand Response Bidding,” in Advanced Computational Methods in Energy, Power, Electric Vehicles, and Their Integration, vol. 763, Singapore: Springer Singapore, 2017, pp. 111–125.
[9] J. Zhu, Optimization of power system operation, Second edition.. Piscataway NJ: Hoboken, New Jersey: IEEE Press ; Wiley, 2015.
[10] Y. V. Makarov, P. V. Etingov, J. Ma, Z. Huang, and K. Subbarao, “Incorporating Uncertainty of Wind Power Generation Forecast Into Power System Operation, Dispatch, and Unit Commitment Procedures,” IEEE transactions on sustainable energy, vol. 2, no. 4, pp. 433–442, 2011, doi: 10.1109/TSTE.2011.2159254.
[11] J. C. Passelergue et al., Market-based real-time dispatch in the hellenic power system. Piscataway NJ: IEEE, 2004.
[12] T. Zhao, X. Pan, S. Yao, C. Ju, and L. Li, “Strategic Bidding of Hybrid AC/DC Microgrid Embedded Energy Hubs: A Two-Stage Chance Constrained Stochastic Programming Approach,” IEEE transactions on sustainable energy, vol. 11, no. 1, pp. 116–125, 2020, doi: 10.1109/TSTE.2018.2884997.
[13] S. A. Hosseini, J.-F. Toubeau, Z. De Grève, and F. Vallée, “An advanced day-ahead bidding strategy for wind power producers considering confidence level on the real-time reserve provision,” Applied energy, vol. 280, 2020, doi: 10.1016/j.apenergy.2020.115973.
[14] H. M. . Pousinho, J. Contreras, P. Pinson, and V. M. . Mendes, “Robust optimisation for self-scheduling and bidding strategies of hybrid CSP–fossil power plants,” International journal of electrical power & energy systems, vol. 67, pp. 639–650, 2015, doi: 10.1016/j.ijepes.2014.12.052.
[15] L. Baringo and R. Sánchez Amaro, “A stochastic robust optimization approach for the bidding strategy of an electric vehicle aggregator,” Electric power systems research, vol. 146, pp. 362–370, 2017, doi: 10.1016/j.epsr.2017.02.004.
[16] X. Han and G. Hug, “A distributionally robust bidding strategy for a wind-storage aggregator,” Electric power systems research, vol. 189, p. 106745, 2020, doi: 10.1016/j.epsr.2020.106745.
[88] Byung-Gook Kim, Yu Zhang, M. van der Schaar, and Jang-Won Lee, “Dynamic Pricing and Energy Consumption Scheduling With Reinforcement Learning,” IEEE transactions on smart grid, vol. 7, no. 5, pp. 2187–2198, 2016, doi: 10.1109/TSG.2015.2495145.

[89] R. Lincoln, S. Galloway, B. Stephen, and G. Burt, “Comparing Policy Gradient and Value Function Based Reinforcement Learning Methods in Simulated Electrical Power Trade,” IEEE transactions on power systems, vol. 27, no. 1, pp. 373–380, 2012, doi: 10.1109/TPWRS.2011.2166091.

[90] S. S. Reka, P. Venugopal, H. H. Alhelou, P. Siano, and M. E. H. Golshan, “Real Time Demand Response Modeling for Residential Consumers in Smart Grid Considering Renewable Energy With Deep Learning Approach,” IEEE access, vol. 9, pp. 56551–56562, 2021, doi: 10.1109/ACCESS.2021.3071993.

[91] X. Zhang, R. Lu, J. Jiang, S. H. Hong, and W. S. Song, “Testbed implementation of reinforcement learning-based demand response energy system management,” Applied energy, vol. 297, p. 117131, 2021, doi: 10.1016/j.apenergy.2021.117131.

[92] Z. Qin, D. Liu, H. Hua and J. Cao, "Privacy Preserving Load Control of Residential Microgrid via Deep Reinforcement Learning,” IEEE Transactions on Smart Grid, vol. 12, no. 5, pp. 4079-4089, Sept. 2021, doi: 10.1109/TSG.2021.3088290.

[93] S. Aladdin, S. El-Tantawy, M. M. Fouda, and A. S. Tag Eldien, “MARLA-SG: Multi-Agent Reinforcement Learning Algorithm for Efficient Demand Response in Smart Grid,” IEEE access, vol. 8, pp. 210626–210639, 2020, doi: 10.1109/ACCESS.2020.3038863.

[94] B. Wang, Y. Li, W. Ming, and S. Wang, “Deep Reinforcement Learning Method for Demand Response Management of Interruptible Load,” IEEE transactions on smart grid, vol. 11, no. 4, pp. 3146–3155, 2020, doi: 10.1109/TSG.2020.2967430.

[95] F. Alfaverh, M. Denai, and Y. Sun, “Demand Response Strategy Based on Reinforcement Learning and Fuzzy Reasoning for Home Energy Management,” IEEE access, vol. 8, pp. 39310–39321, 2020, doi: 10.1109/ACCESS.2020.2974286.

[96] L. Wen, K. Zhou, J. Li, and S. Wang, “Modified deep learning and reinforcement learning for an incentive-based demand response model,” Energy (Oxford), vol. 205, p. 118019, 2020, doi: 10.1016/j.energy.2020.118019.

[97] R. Lu, Y.-C. Li, Y. Li, J. Jiang, and Y. Ding, “Multi-agent deep reinforcement learning based demand response for discrete manufacturing systems energy management,” Applied energy, vol. 276, p. 115473, 2020, doi: 10.1016/j.apenergy.2020.115473.

[98] X. Kong, D. Kong, J. Yao, L. Bai, and J. Xiao, “Online pricing of demand response based on long short-term memory and reinforcement learning,” Applied energy, vol. 271, p. 114945, 2020, doi: 10.1016/j.apenergy.2020.114945.

[99] M. Babar, P. H. Nguyen, V. Cuk, I. G. Kamphuis, M. Bongaerts, and Z. Hanzelka, “The Evaluation of Agile Demand Response: An Applied Methodology,” IEEE transactions on smart grid, vol. 9, no. 6, pp. 6118–6127, 2018, doi: 10.1109/TSG.2017.2703643.

[100] R. Lu, S. H. Hong, and X. Zhang, “A Dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach,” Applied energy, vol. 220, pp. 220–230, 2018, doi: 10.1016/j.apenergy.2018.03.072.

[101] S. Bahrami, V. W. S. Wong, and J. Huang, “An Online Learning Algorithm for Demand Response in Smart Grid,” IEEE transactions on smart grid, vol. 9, no. 5, pp. 4712–4725, 2018, doi: 10.1109/TSG.2017.2667599.

[102] Z. Wen, D. O’Neill, and H. Mael, “Optimal Demand Response Using Device-Based Reinforcement Learning,” IEEE transactions on smart grid, vol. 6, no. 5, pp. 2312–2324, 2015, doi: 10.1109/TSG.2015.2396993.

[103] X. Li, X. Han, and M. Yang, “Day-ahead Optimal Dispatch Strategy for Active Distribution Network Based on Improved Deep Reinforcement Learning,” IEEE access, pp. 1–1, 2022, doi: 10.1109/ACCESS.2022.3141824.

[104] C. Guo, X. Wang, Y. Zheng, and F. Zhang, “Real-time Optimal Energy Management of Microgrid with Uncertainties based on Deep Reinforcement Learning,” Energy (Oxford), vol. 238, p. 121873, 2022, doi: 10.1016/j.energy.2021.121873.

[105] V. H. Bui and W. Su, “Real-time operation of distribution network: A deep reinforcement learning-based reconfiguration approach,” Sustainable energy technologies and assessments, vol. 50, p. 101841, 2022, doi: 10.1016/j.seta.2021.101841.

[106] J. Wang, C. Guo, C. Yu, and Y. Liang, “Virtual power plant containing electric vehicles scheduling strategies based on deep reinforcement learning,” Electric power systems research, vol. 205, p. 107714, 2022, doi: 10.1016/j.enerpr.2021.107714.

[107] J. Li, J. Yao, T. Yu, X. Zhang, “Distributed deep reinforcement learning for integrated generation-control and power-dispatch of interconnected power grid with various renewable units”, pp. 1-20, 2021, doi: 10.1049/rpg2.12310.

[108] T. Yang, L. Zhao, W. Li, and A. Y. Zomaya, “Dynamic energy dispatch strategy for integrated energy system based on improved deep reinforcement learning,” Energy (Oxford), vol. 235, p. 121377, 2021, doi: 10.1016/j.energy.2021.121377.
[109] Q. Li et al., “Integrating Reinforcement Learning and Optimal Power Dispatch to Enhance Power Grid Resilience,” IEEE transactions on circuits and systems, II, Express briefs, pp. 1–1, 2021, doi: 10.1109/TCSII.2021.3131316.

[110] F. Meng, Y. Bai, and J. Jin, “An advanced real-time dispatching strategy for a distributed energy system based on the reinforcement learning algorithm,” Renewable energy, vol. 178, pp. 13–24, 2021, doi: 10.1016/j.renene.2021.06.032.

[111] H. Tang, S. Wang, K. Chang, and J. Guan, “Intra-day Dynamic Optimal Dispatch for Power System Based on Deep Q-Learning,” IEEE transactions on electrical and electronic engineering, vol. 16, no. 7, pp. 954–964, 2021, doi: 10.1002/tee.23379.

[112] J. Li, T. Yu, X. Zhang, F. Li, D. Lin, and H. Zhu, “Efficient experience replay based deep deterministic policy gradient for AGC dispatch in integrated energy system,” Applied energy, vol. 285, p. 116386, 2021, doi: 10.1016/j.apenergy.2020.116386.

[113] X. Sun and J. Qiu, “Two-Stage Volt/Var Control in Active Distribution Networks With Multi-Agent Deep Reinforcement Learning Method,” IEEE transactions on smart grid, vol. 12, no. 4, pp. 2903–2912, 2021, doi: 10.1109/TSG.2021.3052998.

[114] N. Yang et al, “Hierarchical Multi-Agent Deep Reinforcement Learning for Multi-Objective Dispatching in Smart Grid”, China Automation Congress, 2021.

[115] C. Guo, X. Wang, Y. Zheng, and F. Zhang, “Optimal energy management of multi-microgrids connected to distribution system based on deep reinforcement learning,” International journal of electrical power & energy systems, vol. 131, p. 107048, 2021, doi: 10.1016/j.ijepes.2020.110748.

[116] L. Lei, Y. Tan, G. Dahlenburg, W. Xiang, and K. Zheng, “Dynamic Energy Dispatch Based on Deep Reinforcement Learning in IoT-Driven Smart Isolated Microgrids,” IEEE internet of things journal, vol. 8, no. 10, pp. 7938–7953, 2021, doi: 10.1109/IJOT.2020.3042007.

[117] T. Visutarrrom, T. C. Chiang, A. Konak and S. Kulturel-Konak, "Reward Learning-Based Differential Evolution for Solving Economic Dispatch Problems," 2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2020, pp. 913-917, doi: 10.1109/IEEM45057.2020.9309983.

[118] D. Fang, X. Guan, B. Hu, Y. Peng, M. Chen, and K. Hwang, “Deep Reinforcement Learning for Scenario-Based Robust Economic Dispatch Strategy in Internet of Energy,” IEEE internet of things journal, vol. 8, no. 12, pp. 9654–9663, 2021, doi: 10.1109/IJOT.2020.3040294.

[119] T. A. Nakabi and P. Toivanen, “Deep reinforcement learning for energy management in a microgrid with flexible demand,” Sustainable Energy, Grids and Networks, vol. 25, p. 100413, 2021, doi: 10.1016/j.segan.2020.100413.

[120] G. Zhang et al., “Data-driven optimal energy management for a wind-solar-diesel-battery-reverse osmosis hybrid energy system using a deep reinforcement learning approach,” Energy conversion and management, vol. 227, 2021, doi: 10.1016/j.enconman.2020.113608.

[121] M. Biemann, X. Liu, Y. Zeng, and L. Huang, “Addressing partial observability in reinforcement learning for energy management,” in BuildSys 2021 - Proceedings of the 2021 ACM International Conference on Systems for Energy-Efficient Built Environments, 2021, pp. 324–328, doi: 10.1145/3486611.3488730.

[122] K. Lv, H. Tang, B. Bak-Jensen, J. Radhakrishna Pillai, Q. Tan, Q. Zhang, “Hierarchical learning optimisation method for the coordination dispatch of the inter-regional power grid considering the quality of service index”, IET Generation, Transmission & Distribution, 2020, 14, (18), p. 3673-3684, doi: 10.1049/iet-gtd.2019.1869

[123] S. Wu, W. Hu, Z. Lu, Y. Gu, B. Tian, and H. Li, “Power System Flow Adjustment and Sample Generation Based on Deep Reinforcement Learning,” Journal of modern power systems and clean energy, vol. 8, no. 6, pp. 1115–1127, 2020, doi: 10.35833/MPCE.2020.000240.

[124] Z. Yan and Y. Xu, “Real-Time Optimal Power Flow: A Lagrangian Based Deep Reinforcement Learning Approach,” IEEE transactions on power systems, vol. 35, no. 4, pp. 3270–3273, 2020, doi: 10.1109/TPWRDS.2020.2987292.

[125] Z. Deng et al., “Coordinated Optimization of Generation and Compensation to Enhance Short-Term Voltage Security of Power Systems Using Accelerated Multi-Objective Reinforcement Learning,” IEEE access, vol. 8, pp. 34770–34782, 2020, doi: 10.1109/ACCESS.2020.2974503.

[126] S. H. Oh, Y. T. Yoon, and S. W. Kim, “Online reconfiguration scheme of self-sufficient distribution network based on a reinforcement learning approach,” Applied energy, vol. 280, 2020, doi: 10.1016/j.apenergy.2020.115900.

[127] Y. Shang et al., “Stochastic dispatch of energy storage in microgrids: An augmented reinforcement learning approach,” Applied energy, vol. 261, p. 114423, 2020, doi: 10.1016/j.apenergy.2019.114423.

[128] H. Wang, Z. Lei, X. Zhang, J. Peng, and H. Jiang, “Multiobjective Reinforcement Learning-Based Intelligent Approach for Optimization of Activation Rules in Automatic Generation Control,” IEEE access, vol. 7, pp. 17480–17492, 2019, doi: 10.1109/ACCESS.2019.2894756.

[129] L. Ya, Z. Deliang and W. Xuanyuan, “A Peak Regulation Ancillary Service Optimal Dispatch Method of Virtual Power Plant Based on Reinforcement Learning,” 2019 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia), 2019, pp. 4356-4361, doi: 10.1109/ISGT-Asia.2019.8881083.
[130] H. Hua, Y. Qin, C. Hao, and J. Cao, “Optimal energy management strategies for energy Internet via deep reinforcement learning approach,” Applied energy, vol. 239, pp. 598–609, 2019, doi: 10.1016/j.apenergy.2019.01.145.

[131] Y. Ji, J. Wang, J. Xu, X. Fang, and H. Zhang, “Real-time energy management of a microgrid using deep reinforcement learning,” Energies (Basel), vol. 12, no. 12, p. 2291, 2019, doi: 10.3390/en12122291.

[132] Y. Du and F. Li, “Intelligent Multi-Microgrid Energy Management Based on Deep Neural Network and Model-Free Reinforcement Learning,” IEEE transactions on smart grid, vol. 11, no. 2, pp. 1066–1076, 2020, doi: 10.1109/TSG.2019.2930299.

[133] J. Duan et al., “A Deep Reinforcement Learning Based Approach for Optimal Active Power Dispatch,” 2019.

[134] C. Han, B. Yang, T. Bao, T. Yu, and X. Zhang, “Bacteria foraging reinforcement learning for risk-based economic dispatch via knowledge transfer,” Energies (Basel), vol. 10, no. 5, p. 638, 2017, doi: 10.3390/en10050638.

[135] B. V. Mbuwir, F. Ruelens, F. Spiessens, and G. Deconinck, “Battery energy management in a microgrid using batch reinforcement learning,” Energies (Basel), vol. 10, no. 11, p. 1846, 2017, doi: 10.3390/en10111846.

[136] P. Kofinas, G. Vouros, and A. I. Dounis, “Energy management in solar microgrid via reinforcement learning using fuzzy reward,” Advances in building energy research, vol. 12, no. 1, pp. 97–115, 2018, doi: 10.1080/17512549.2017.1314832.

[137] E. Kuznetsova, Y.-F. Li, C. Ruiz, E. Zio, G. Ault, and K. Bell, “Reinforcement learning for microgrid energy management,” Energy (Oxford), vol. 59, pp. 133–146, 2013, doi: 10.1016/j.energy.2013.05.060.

[138] L. Xiong et al., “A Two-Level Energy Management Strategy for Multi-Microgrid Systems With Interval Prediction and Reinforcement Learning,” IEEE transactions on circuits and systems. I, Regular papers, pp. 1–12, 2022, doi: 10.1109/TCSI.2022.3141229.

[139] M. Al-Saffar and P. Musilek, “Distributed Optimization for Distribution Grids With Stochastic DER Using Multi-Agent Deep Reinforcement Learning,” IEEE access, vol. 9, pp. 63059–63072, 2021, doi: 10.1109/ACCESS.2021.3075247.

[140] D. Mao, L. Ding, C. Zhang, H. Rao and G. Yan, "Multi-Agent Reinforcement Learning-based Distributed Economic Dispatch Considering Network attacks and Uncertain Costs," 2021 IEEE 16th Conference on Industrial Electronics and Applications (ICIEA), 2021, pp. 469-474, doi: 10.1109/ICIEA51954.2021.9516143.

[141] D. Li, L. Yu, N. Li, and F. Lewis, “Virtual-Action-Based Coordinated Reinforcement Learning for Distributed Economic Dispatch,” IEEE transactions on power systems, vol. 36, no. 6, pp. 5143–5152, 2021, doi: 10.1109/TPWRS.2021.3070161.

[142] X. Fang, Q. Zhao, J. Wang, Y. Han, and Y. Li, “Multi-agent Deep Reinforcement Learning for Distributed Energy Management and Strategy Optimization of Microgrid Market,” Sustainable cities and society, vol. 74, p. 103163, 2021, doi: 10.1016/j.scs.2021.103163.

[143] R. Hao, T. Lu, Q. Ai and H. He, "Distributed Online Dispatch For Microgrids Using Hierarchical Reinforcement Learning Embedded With Operation Knowledge," in IEEE Transactions on Power Systems, doi: 10.1109/TPWRS.2021.3092220.

[144] E. Foruzan, L.-K. Soh, and S. Asgarpoor, “Reinforcement Learning Approach for Optimal Distributed Energy Management in a Microgrid,” IEEE transactions on power systems, vol. 33, no. 5, pp. 5749–5758, 2018, doi: 10.1109/TPWRS.2018.2823641.

[145] S. Stanko and K. Macek, “Risk-averse Distributional Reinforcement Learning: A CVaR Optimization Approach”, 11th International Conference on Neural Computation Theory and Applications, 2019, doi: 10.5220/0008175604120423.

[146] M. Godbout, M. Heuillet, S. Chandra, R. Bhati, and A. Durand, “CARL: Conditional-value-at-risk Adversarial Reinforcement Learning,” arXiv preprint arXiv: 2109.09470.

[147] Y. Gao, K. Y. C. Lui, and P. Hernandez-Leal, “Robust Risk-Sensitive Reinforcement Learning Agents for Trading Markets,” arXiv preprint arXiv: 2107.08083.

[148] Y. J. Ma, D. Jayaraman, and O. Bastani, “Conservative Offline Distributional Reinforcement Learning,” arXiv preprint arXiv: 2107.06106

[149] E. Vittori, M. Trapletti, and M. Restelli, “Option Hedging with Risk Averse Reinforcement Learning,” arXiv preprint arXiv: 2010.12245.

[150] R. Singh, Q. Zhang, and Y. Chen, “Improving Robustness via Risk Averse Distributional Reinforcement Learning,” arXiv preprint arXiv: 2005.00585.

[151] X. Ma, L. Xia, Z. Zhou, J. Yang, and Q. Zhao, “DSAC: Distributional Soft Actor Critic for Risk-Sensitive Reinforcement Learning,” arXiv preprint arXiv: 2004.14547.

[152] B. Zhou, H. Zeng, F. Wang, Y. Li, and H. Tian, “Efficient and Robust Reinforcement Learning with Uncertainty-based Value Expansion,” arXiv preprint arXiv: 1912.05328.
[153] P. Doshi and P. J. Gmytrasiewicz, “Monte carlo sampling methods for approximating interactive POMDPs,” The Journal of artificial intelligence research, vol. 34, pp. 297–337, 2009, doi: 10.1613/jair.2630.

[154] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training Recurrent Neural Networks,” International conference on machine learning, pp. 1310-1318, PMLR, 2013.

[155] Y. Yu, X. Si, C. Hu, and J. Zhang, “A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures,” Neural computation, vol. 31, no. 7, pp. 1235–1270, 2019, doi: 10.1162/neco_a_01199.

[156] A. Pillay, S. Prabhakar Karthikeyan, and D. Kothari, “Congestion management in power systems – A review,” International journal of electrical power & energy systems, vol. 70, pp. 83–90, 2015, doi: 10.1016/j.ijepes.2015.01.022.

[157] G. G. Fiuza de Bragança and T. Daglish, “Can market power in the electricity spot market translate into market power in the hedge market?,” Energy economics, vol. 58, pp. 11–26, 2016, doi: 10.1016/j.eneco.2016.05.010.

[158] N. C. Dormady, D. Mazmanian, A. Z. Rose, S. Wilkie, J. Jurewitz, and University of Southern California. Policy, Planning Development, Emissions markets, power markets and market power: A study of the interactions between contemporary emissions markets and deregulated power markets. 2012.

[159] Z. Zhu, K. W. Chan, S. Bu, S. W. Or, X. Gao, and S. Xia, “Analysis of Evolutionary Dynamics for Bidding Strategy Driven by Multi-Agent Reinforcement Learning,” IEEE transactions on power systems, vol. 36, no. 6, pp. 5975–5978, 2021, doi: 10.1109/TPWRS.2021.3099693.

[160] Y. Zhao et al., “Local Differential Privacy-Based Federated Learning for Internet of Things,” in IEEE Internet of Things Journal, vol. 8, no. 11, pp. 8836-8853, 1 June 1, 2021, doi: 10.1109/JIOT.2020.3037194.

[161] H. Iiduka, “Appropriate Learning Rates of Adaptive Learning Rate Optimization Algorithms for Training Deep Neural Networks,” in IEEE Transactions on Cybernetics, doi: 10.1109/TCYB.2021.3107415.

[162] A. Nguyen, J. Yosinski, and J. Clune, “Understanding Neural Networks via Feature Visualization: A Survey,” in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Cham: Springer International Publishing, 2019, pp. 55–76.

[163] A. Marot et al., “Learning to run a power network challenge for training topology controllers,” Electric power systems research, vol. 189, p. 106635, 2020, doi: 10.1016/j.epsr.2020.106635.

[164] H. Chen, Stochastic approximation and its applications. Dordrecht: Kluwer Academic Publishers, 2002.

[165] P. Abbeel and A. Ng, “Apprenticeship learning via inverse reinforcement learning,” in Proceedings of the twenty-first international conference on machine learning, 2004, pp. 1–8, doi: 10.1145/1015330.1015430.