Deep Reinforcement Learning in Power Distribution Systems: Overview, Challenges, and Opportunities

Yuanqi Gao, *Member, IEEE,* Nanpeng Yu, *Senior Member, IEEE,*

***Abstract*—To facilitate the integration of distributed energy resources and improve existing operational strategies, power distribution systems have seen a rapid proliferation of deep reinforcement learning (DRL) based applications. DRL approach is well suited for dynamic, complex, and uncertain operational environments such as power distribution systems. This paper reviews the rapidly growing body of literature that develops applications of reinforcement learning in power distribution systems. These applications include active grid management, energy management system, retail electricity market, and demand response. This paper also summarizes the challenges of deploying DRL based solutions in distribution systems such as safety, robustness, interpretability, and sample efficiency. Finally, the research opportunities that can be pursued to address the challenges are provided.**

***Index Terms*—Power distribution systems, deep reinforcement learning.**

I. Introduction

Over the last two decades, power distribution systems have experienced tremendous change with the penetration of distributed energy resources (DERs) such as electric vehicles, behind-the-meter solar photovoltaic systems, and battery storage devices. These DERs brought significant operational challenges including dramatic voltage deviations, bidirectional power flow, and deteriorating power quality. To overcome the operational challenges in distribution systems, a considerable amount of infrastructure such as advanced metering infrastructure and remotely controllable devices such as tie switches, on load tap changers, and smart inverters have been installed.

To fully utilize the advanced equipment and coordinate the operations of DERs, many physical model based control and optimization algorithms were developed and adopted by distribution system operators. However, most of these physical model based control algorithms are built based on the distribution network topology, parameter, and customer information in the geographical information system and customer management system. However, these methods did not achieve satisfactory results in practice. This is because electric utilities often do not have reliable and accurate information about the distribution network or the end-use customers. Furthermore, the computation time of the physical model based algorithms often increases quickly with the problem size. This makes them unfitted for real-time operations.

To mitigate the problems of distribution model uncertainty and computational complexity, an alternative approach is to learn control and optimization policies based on the multitude of data generated from the real-world distribution system or a high fidelity simulation environment. Reinforcement learning (RL) algorithm is one of the most promising data-driven approaches to solve sequential decision making problems. RL aims at learning optimal control policies in Markov decision processes (MDPs) [1]. In the RL setup, an agent interacts with an environment by observing the environment states, executing actions, and receiving numerical rewards. The goal of the agent is to learn a policy, which is a mapping from states to actions, such that the expected total discounted reward is maximized. For MDPs having a finite number of states and actions, policy learning can be achieved via discrete stochastic dynamic programming or policy iteration methods. To scale RL algorithms to MDPs with large state and action domains, deep neural network parameterized policy and value functions have been adopted by a large body of work in Deep RL (DRL).

DRL algorithms can be classified in a number of ways. First, DRL algorithms can be grouped into on-policy and off-policy algorithms, which differ about whether or not to evaluate and improve a different policy than the current one. Second, DRL algorithms can be categorized into value based and policy gradient/actor critic based on whether a policy function is explicitly maintained. Third, the RL setup could be either online or in a batch model. The online RL setup integrates the data collection and the learning process. In the Batch RL setup, the data collection is decoupled from the learning process. In other words, the RL agent learns from a fixed set of experiences. Finally, we have model-free RL and model-based DRL algorithms. The model-free RL agents directly learn the value or policy function from data, whereas model-based algorithms build a model for the environment transition function for learning and planning.

Although DRL has achieved great success in many domains such as Atari games [2], data center cooling system control [3], and Go game [4], it has not seen wide-spread adoption in critical infrastructure systems such as power distribution systems because they must be operated in an extremely reliable manner. In addition, the real-world operation environment is highly complex with very high dimensional state and action space. The unreliable and time varying distribution system model and intermittent DERs make it even harder to achieve satisfactory results with DRL based approach. To achieve the ambitious goal of successfully deploying DRL based solutions to power distribution systems, further research and developments are needed. This paper aims at identifying the challenges of DRL based solutions for distribution system applications and providing ideas for future research directions.

There are several papers that provide a high level review

Y. Gao and N. Yu are with the Department of Electrical and Computer Engineering, University of California, Riverside, CA, 92521.

of applications of DRL in power systems [5], [6], [7]. In comparison, this paper focuses on control-related topics in power distribution systems. Moreover, this paper delivers in-depth analysis of the challenges and research opportunities for designing DRL algorithms that could be eventually deployed in operational environments of power distribution systems.

This paper has two contributions. First, it presents an up-to-date literature review on deep reinforcement learning in power distribution systems. Second, it summarizes the challenges and research opportunities of DRL in power distribution systems.

The remainder of this paper is organized as follows. Section II reviews applications of DRL in power distribution systems. Section III discusses the challenges and opportunities. Section IV provides the concluding remarks.

II. Applications of RL in Distribution Systems

This section reviews the promising applications of DRL in power distribution systems. We group the existing literature into the following application areas: active distribution grid management, energy management system and retail electricity market, and demand response. Due to the space limitation, only selected papers are reviewed. A summary of RL methodologies in the literature is presented at the end of this section.

A. Active Distribution Grid Management

1) Volt-VAR Control (VVC): There are a large number of papers that tackle the VVC problem using RL methods. An early work proposed the use of Q-learning for reactive power control [8]. The actions include transformer taps, shunt compensations, voltage and power generations at PV buses. The reward is proportional to the degree of operating limit satisfaction. To reduce the communication and computation burden of a central controller, [9] proposed a multi-agent Q-learning VVC framework. The actions are generator bus voltage, capacitor banks, and tap positions of transformers.

To handle high-dimensional and continuous state spaces, RL algorithms have been combined with function approximators. [10] proposed a least square policy iteration (LSPI) algorithm for tap changer control, where the states are the nodal voltages and existing tap positions. A linear function approximation of the Q-function is constructed and updated using the standard LSPI iteration. [11] formulates a multi-agent MDP for the VVC problem and proposes a multi-agent deep Q learning algorithm. The action space is decomposed in a per-device manner to improve the learning efficiency. [12] developed a consensus multi-agent RL algorithm in the maximum entropy RL framework, and demonstrated the algorithm's robustness against agent/communication failures. While the operation constraints were handled by domain knowledge based reward designs in the aforementioned literature, they have also been rigorously modeled in the constrained Markov decision process (CMDP) framework. In [13], the VVC problem is formulated as a CMDP, where the voltage violations are treated as constraints. This paper adopts the constrained policy optimization (CPO) algorithm to solve the CMDP problem. To further improve the sampling efficiency of the DRL algorithm, [14] developed a safe and off-policy RL algorithm called the constrained soft actor-critic (CSAC) to solve the VVC problems.

RL has also been adopted to perform smart inverter control for VVC problems. [15] proposed to use the deep deterministic policy gradient (DDPG) to improve voltage profile and minimize the curtailment of solar PV generation. To coordinate devices operating in different time scales, [16] proposed a two-time scale VVC for joint smart inverter and capacitor control. The fast time-scale smart inverters control problem is solved by quadratic programming and the slow time-scale capacitors control problem is conducted by a DQN agent.

2) Distribution Network Reconfiguration and Restoration: Researchers have applied RL for the system restoration or the network reconfiguration process. [17] proposed a hybrid multi-agent Q-learning algorithm to determine the opening/closing switches for fast system restoration. [18] proposed a shipboard power system reconfiguration algorithm based on Q-learning. The paper focused on finding the best sequence of open/close pairs to reach a final static configuration which takes the shortest amount of execution time. [19] proposed a multi-agent Q-learning based distribution system restoration framework for load restoration. Similarly, [20] proposed a Q-learning framework for distribution network service restoration and load management. Switching device status optimization can also reduce network losses. Early work adopted tabular Q-learning algorithm for static minimum loss network reconfiguration [21]. The actions are to change the switch status. To avoid specifying an accurate simulation model or learning by interacting with the grid, batch RL methods have been developed to solve the reconfiguration problem. [22] proposed a DQN based approach to learn from the historical dataset and a synthetic dataset generated from a Gaussian process to improve the training data diversity. The problem of bias for learning from a fixed dataset has also been addressed by regularizing the learned policy toward the behavior policy [23].

B. Energy Management System and Retail Electricity Market

A number of papers have applied RL for the residential and microgrid energy management problem. In [24], the battery storage operation problem is formulated as an MDP and solved with Q-learning considering three possible charging/discharging actions. [25] extended the framework by adopting the proximal policy optimization (PPO) algorithm with a recurrent neural network to represent the price time series. [26] proposed a prioritized DDPG based residential multi-energy system. The learned policy is capable of determining power dispatching signals for distributed generation and energy storage systems (ESS) that minimizes the microgrid operational cost while satisfying operational constraints. A bi-level microgrid management framework is developed in [27]. The first-level control is carried out by a Q-learning agent, which learns an optimal location price control policy. In the second level, each microgrid manages its own generation/storage by solving a mixed-integer nonlinear programming problem.

The RL-based retail electricity market was also explored by researchers. [28] introduced a customer-centric model for the local event-driven market (LEM), where the RL agent tries to

learn a market clearing strategy. The problem is solved by a modified Q-learning accounting for different rewards across episodes. [29] proposed a microgrid trading framework to determine the energy trading strategy with other microgrids that increases its profit.

C. Demand Response

[30] proposed a modified fitted Q-iteration for demand response using thermostatically controlled loads. The state variables are divided into controllable parts (e.g., indoor temperature) and non-controllable parts (e.g., outdoor temperature). The action is temperature control which is proposed by an RL agent but rectified by a backup controller to guarantee safety and comfort. An RL-based solution is introduced for multi-user demand response considering real-time pricing (RTP) [31]. Since the price for each user is influenced by the consumption of all users, the DR problem is formulated as a partially observable Markov game (POMG), in which each user controls the sleep/awake status of its loads to minimize the electricity bill and discomfort while participating in the DR program. The POMG problem is solved by estimating other users' observations to form an estimated fully observable MG, which is solved by an actor-critic algorithm.

D. Summary of RL-based Applications in Distribution Grid

The existing literature leveraged a wide range of RL algorithms in distribution system applications. Early algorithms mostly utilized tabular Q-learning for problems with finite and low dimensional state space. Initially, problems with larger state spaces are often addressed by heuristics that augment the Q-table. To process more complex state spaces, the researchers eventually adopt DRL algorithms that leverage deep neural networks. Off-the-shelf DRL approaches such as action value methods (e.g., DQN and DDQN), policy gradient methods (e.g., PPO, and CPO), and actor critic framework (e.g., DDPG and SAC) have been explored. In addition to the standard DRL algorithms, innovative ones such as CSAC and batch constrained SAC (BCSAC) [23] have been developed to address the technical issues associated with operational safety and limited historical training samples.

In terms of the RL environment setup, some literature assumes that RL agents learn from a set of historical data. This setup avoids specifying a complete and accurate distribution system simulation model, which the operators most likely do not have access to. In other papers, the RL agents interact with a simulated distribution network. Learning by directly interacting with the physical environment is also explored. To prevent the selection of risk actions, safety checks by simulation models or human operators are often required.

The mathematical formulation of the RL problem also varies in the existing literature. In most of the papers, the RL problem is formulated as an MDP or a CMDP. In a few papers, the actions affect only the immediate rewards but not future states. In this case, the agent is facing a multi-armed bandit problem rather than a regular MDP. In some problem formulation, the learning environment is not strictly Markovian. However, by aggregating system information from multiple time steps to form the state vector, DRL algorithms can still be applied.

III. DRL in Distribution Systems: Challenges and Opportunities

In this section, we discuss the main challenges in the development and deployment of DRL algorithms in distribution systems. Several research and development opportunities are pointed out to address the technical challenges.

A. Challenges

1) Safety and Robustness: To operate a power distribution grid with many critical infrastructure in a safe and reliable manner, the learned control policy must be robust to measurement noise and unforeseen operation conditions such as addition of distribution generation or change in network topology. Moreover, critical distribution system operational constraints must be satisfied all the time. Relying on backup controllers or human operators to override the RL agent's control signals in real-time is not an ideal operational risk mitigation strategy. Although existing literature already leveraged the constrained MDP framework to develop safe RL algorithms for applications such as Volt-VAR optimization, this type of safe RL algorithms could only guarantee near but not strict operational constraint satisfaction.

2) Interpretability of DRL policies: Despite promising performance of DRL-based algorithms in distribution system applications, the learned control policies are embedded in deep neural networks. This makes it difficult to interpret the learned policy, explain them to the system operators, and check for desired safety properties. In order to improve the user acceptance of DRL in power distribution systems, we must develop DRL algorithms that are interpretable. For example, the policy must be presented in a format that allows electric utilities to evaluate the voltage margins under different operational conditions.

3) Sample efficiency: To learn a good control policy, DRL algorithms typically require a large number of training data. This is problematic when the training data can only be sampled from historical operational experiences in the real-world distribution system. Excessive exploration in the real-world system degrades the algorithm performance in the short term when a good policy can not be obtained on time to compensate for the cost due to trial-and-error. Off-policy RL algorithms offer a natural way to improve the sample efficiency by reusing previously collected data. However, designing off-policy RL algorithms to learn accurate value functions or near optimal policies for large-scale distribution system applications is still a very challenging problem. Although batch DRL algorithms can infer good control policies from existing operational data without exploration. The learned policies are not guaranteed to be near optimal. This is especially troublesome when the utilities do not have abundant and diverse historical data.

4) Availability of accurate training environments: Most of the electric utilities do not have accurate and reliable network topology documents, network parameter estimates, and load and DG models. These practical challenges make it difficult to build a trustworthy distribution system simulation environment for training RL algorithms. Although many data-driven network topology, parameter estimation and load modeling

algorithms have been developed, almost none of them could guarantee extremely small estimation error for a wide range of distribution systems. Therefore, RL agents must be able to transfer the knowledge they learned in the simulation environment to the real-world distribution systems. It has been shown that naive continuous training on new data streams does not yield satisfactory learning results. Worse yet, re-training the RL agent in a different environment from the previous checkpoints still requires a large amount of new samples.

5) Lack of standardized test cases, testing procedure, and performance metrics: One of the key drivers for major advances in machine learning research is the availability of high quality standard training and testing dataset. However, to date no standardized test cases and dataset are specifically designed for RL research and development in the area of power distribution systems. This makes it extremely difficult to conduct a fair comparison between a new RL algorithm and the benchmark for power distribution system applications. In addition, the testing procedure and performance metrics reporting used in the power system literature are not always carried out in an appropriate manner. Furthermore, many RL implementation details such as hyperparameters, random seeds, and software codes are often missing from the literature in power distribution systems. This makes it difficult to verify and reproduce RL research results.

6) Non-Markovian environment: The RL problems are formulated within the Markov decision process framework. However, the Markovian property may not be valid for some of the distribution system applications. For example, the time varying price of retail electricity markets and the load patterns are hardly Markov processes. Although RL algorithms such as Q-learning are found to converge for a larger class of problems than MDP, it is difficult to predict if the developed RL algorithm works for a specific non-Markovian distribution system problem domain. The current approaches to deal with Non-Markovian problems are to aggregate the information from the past, such as load/DG time series or price signals, as part of the state. However, these approaches are not often developed in a principled manner.

B. Opportunities

1) Safe exploration: To learn a safe and reliable control policy, the RL agent must satisfy the distribution system operational constraints both during and after the learning phase. In addition, these operational constraints should be enforced strictly rather than on expectation as in the CMDP framework. One may consider modeling the operational constraints in a state-wise manner and adding an action correction layer to guarantee operational constraints satisfaction [32]. An alternative strategy for guaranteeing state-wise constraint satisfaction can be implemented by extracting a verifiable policy, such as decision trees [33], from a given pre-trained policy. Then, we can verify if certain actions will violate operational constraints.

2) Interpretable policy learning: To improve the adoption of and the trust in RL policies, we will need to describe the learned policy to the distribution system operators in the languages familiar to them instead of using descriptions involving neural network terminologies and MDP theory. This can be achieved by programmatic policy learning, which search for and construct policies from a set of domain-specific programming languages [34]. This approach could produce human-readable policies, and allow for verifying the RL agents' behavior for any input state. A policy may also be interpreted by learning structural causal models, which encode causal relationships between variables of interest, such as an action or a reward [35]. These models are then used to generate explanations of the RL agent's behaviors, which are subsets of causes together with the causal chains to the reward variables.

3) Model-based learning and planning: Model-based RL learns a model (state transition probability and reward) then plans a policy within the learned model. They are in general more sample efficient than model-free RL algorithms because the probability distribution model of the environment can be learned separately by supervised learning methods. In addition, prior knowledge about the environment can be specified in the model to improve the learning efficiency. Although inaccurate models can introduce biases in the learned policy, recent model-based policy optimization techniques have shown that such biases can be mitigated by strategically varying the length of model rollouts [36]. It has been demonstrated that these model-based policy optimization techniques greatly improve the model usability. The model-based RL algorithms can also be used in batch RL setup [37].

4) Imitation learning and inverse RL: The operations of power distribution grids usually have several objectives such as operational cost minimization and system resiliency and reliability maximization. It is difficult to strike a balance between various operating objectives in the reward design process. To address this problem, imitation learning and inverse RL [38] can be used to infer useful reward information from a set of demonstrations, which are the historical operational data in power distribution systems.

5) Leveraging specific reward structures: Most off-the-shelf DRL algorithms are designed with very general reward formulations. However, more efficient algorithms can be developed by exploiting specific problem structures. In some distribution system applications, it can be shown that the optimal policy is the greedy policy with respect to the immediate reward. Handcrafted functions can also be selected based on distribution system domain knowledge as the reward. Efficient algorithms such as the online contextual bandit SquareCB [39] and the batch contextual bandit techniques [40] can be leveraged to improve the learning efficiency.

6) Robust RL and transfer learning: RL control policy learned based on operational experiences from one particular distribution system environment may not work well in another one with different topology or set of DERs. To prevent learning from scratch again, one could adopt transfer learning techniques, which transplant a portion of the previously learned policy and value function neural network parameters into the neural networks of the new operating environment. Another approach to deal with uncertain real-world operational environments is to learn a robust policy [41], which maximizes the worst case value function of all the possible MDPs associated with a set of possible real-world operating environments.

IV. Conclusion

This paper provides a literature survey for recent applications of deep reinforcement learning in power distribution systems. We point out that despite the rapid development of the RL algorithms for power distribution systems, there remains six fundamental challenges of adopting RL methods in real-world operations: safety, interpretability, sample efficiency, model uncertainty, non-Markovity, and lack of standardized test cases. Six exciting research and development opportunities for DRL in power distribution systems are identified.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[3] N. Lazic, C. Boutilier, T. Lu, E. Wong, B. Roy, M. Ryu, and G. Imwalle, "Data center cooling using model-predictive control," in *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 3814–3823.

[4] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play," *Science*, vol. 362, no. 6419, pp. 1140–1144, 2018.

[5] Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," *CSEE Journal of Power and Energy Systems*, vol. 6, no. 1, pp. 213–225, 2020.

[6] T. Yang, L. Zhao, W. Li, and A. Y. Zomaya, "Reinforcement learning in sustainable energy and electric systems: A survey," *Annual Reviews in Control*, vol. 49, pp. 145–163, 2020.

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

[8] J. G. Vlachogiannis and N. D. Hatziargyriou, "Reinforcement learning for reactive power control," *IEEE Transactions on Power Systems*, vol. 19, no. 3, pp. 1317–1325, Aug 2004.

[9] Y. Xu, W. Zhang, W. Liu, and F. Ferrese, "Multiagent-based reinforcement learning for optimal reactive power dispatch," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1742–1751, Nov 2012.

[10] H. Xu, A. D. Domínguez-García, and P. W. Sauer, "Optimal tap setting of voltage regulation transformers using batch reinforcement learning," *IEEE Transactions on Power Systems*, vol. 35, no. 3, pp. 1990–2001, 2020.

[11] Y. Zhang, X. Wang, J. Wang, and Y. Zhang, "Deep reinforcement learning based Volt-VAR optimization in smart distribution systems," *arXiv preprint arXiv:2003.03681*, 2020.

[12] Y. Gao, W. Wang, and N. Yu, "Consensus multi-agent reinforcement learning for Volt-VAR control in power distribution networks," *arXiv preprint arXiv:2007.02991*, 2020.

[13] W. Wang, N. Yu, J. Shi, and Y. Gao, "Volt-VAR control in power distribution systems with deep reinforcement learning," in *IEEE SmartGridComm*, 2019, pp. 1–7.

[14] W. Wang, N. Yu, Y. Gao, and J. Shi, "Safe off-policy deep reinforcement learning algorithm for Volt-VAR control in power distribution systems," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3008–3018, 2020.

[15] C. Li, C. Jin, and R. Sharma, "Coordination of PV smart inverters using deep reinforcement learning for grid voltage regulation," *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Dec 2019.

[16] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, "Two-timescale voltage control in distribution grids using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2313–2323, 2020.

[17] D. Ye, M. Zhang, and D. Sutanto, "A hybrid multiagent framework with Q-learning for power grid systems restoration," *IEEE Transactions on Power Systems*, vol. 26, no. 4, pp. 2434–2441, 2011.

[18] S. Das, S. Bose, S. Pal, N. N. Schulz, C. M. Scoglio, and B. Natarajan, "Dynamic reconfiguration of shipboard power systems using reinforcement learning," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 669–676, 2013.

[19] J. Hong, "A multiagent Q-learning-based restoration algorithm for resilient distribution system operation," 2017, Electronic Theses and Dissertations, 2004-2019. 5572. https://stars.library.ucf.edu/etd/5572.

[20] L. R. Ferreira, A. R. Aoki, and G. Lambert-Torres, "A reinforcement learning approach to solve service restoration and load management simultaneously for distribution networks," *IEEE Access*, vol. 7, pp. 145 978–145 987, 2019.

[21] J. G. Vlachogiannis and N. Hatziargyriou, "Reinforcement learning (RL) to optimal reconfiguration of radial distribution system (RDS)," in *Hellenic Conference on Artificial Intelligence*. Springer, 2004, pp. 439–446.

[22] Y. Gao, J. Shi, W. Wang, and N. Yu, "Dynamic distribution network reconfiguration using reinforcement learning," in *IEEE SmartGridComm*, 2019, pp. 1–7.

[23] Y. Gao, W. Wang, J. Shi, and N. Yu, "Batch-constrained reinforcement learning for dynamic distribution network reconfiguration," *IEEE Transactions on Smart Grid*, pp. 1–1, 2020.

[24] H. Wang and B. Zhang, "Energy storage arbitrage in real-time markets via reinforcement learning," in *2018 IEEE Power Energy Society General Meeting (PESGM)*. IEEE, 2018, pp. 1–5.

[25] X. Hanchen, L. Xiao, Z. Xiangyu, and Z. Junbo, "Arbitrage of energy storage in electricity markets with deep reinforcement learning," *arXiv preprint arXiv:1904.12232*, 2019.

[26] Y. Ye, D. Qiu, X. Wu, G. Strbac, and J. Ward, "Model-free real-time autonomous control for a residential multi-energy system using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3068–3082, 2020.

[27] Q. Zhang, K. Dehghanpour, Z. Wang, and Q. Huang, "A learning-based power management method for networked microgrids under incomplete information," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1193–1204, 2020.

[28] T. Chen and W. Su, "Indirect customer-to-customer energy trading with reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 4338–4348, 2019.

[29] X. Lu, X. Xiao, L. Xiao, C. Dai, M. Peng, and H. V. Poor, "Reinforcement learning-based microgrid energy trading with a reduced power plant schedule," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 728–10 737, 2019.

[30] F. Ruelens, B. J. Claessens, S. Vandael, B. De Schutter, R. Babuška, and R. Belmans, "Residential demand response of thermostatically controlled loads using batch reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2149–2159, 2017.

[31] 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.

[32] G. Dalal, K. Dvijotham, M. Vecerik, T. Hester, C. Paduraru, and Y. Tassa, "Safe exploration in continuous action spaces," *arXiv preprint arXiv:1801.08757*, 2018.

[33] O. Bastani, Y. Pu, and A. Solar-Lezama, "Verifiable reinforcement learning via policy extraction," in *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 2494–2504.

[34] A. Verma, V. Murali, R. Singh, P. Kohli, and S. Chaudhuri, "Programmatically interpretable reinforcement learning," *arXiv preprint arXiv:1804.02477*, 2018.

[35] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, "Explainable reinforcement learning through a causal lens," *arXiv preprint arXiv:1905.10958*, 2019.

[36] M. Janner, J. Fu, M. Zhang, and S. Levine, "When to trust your model: Model-based policy optimization," *arXiv preprint arXiv:1906.08253*, 2019.

[37] R. Kidambi, A. Rajeswaran, P. Netrapalli, and T. Joachims, "MOReL: Model-based offline reinforcement learning," *arXiv preprint arXiv:2005.05951*, 2020.

[38] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters, "An algorithmic perspective on imitation learning," *Foundations and Trends in Robotics*, vol. 7, no. 1-2, p. 1–179, 2018.

[39] D. J. Foster and A. Rakhlin, "Beyond UCB: Optimal and efficient contextual bandits with regression oracles," *arXiv preprint arXiv:2002.04926*, 2020.

[40] M. Dudík, J. Langford, and L. Li, "Doubly robust policy evaluation and learning," *arXiv preprint arXiv:1103.4601*, 2011.

[41] M. A. Abdullah, H. Ren, H. B. Ammar, V. Milenkovic, R. Luo, M. Zhang, and J. Wang, "Wasserstein robust reinforcement learning," *arXiv preprint arXiv:1907.13196*, 2019.