Deep Reinforcement Learning for RIS-Aided Full-Duplex Systems: Advances and Challenges

Alice Faisal, *Graduate Member, IEEE,* Ibrahim Al-Nahhal, *Senior Member, IEEE,*
Octavia A. Dobre, *Fellow, IEEE,* Telex M. N. Ngatched, *Senior Member, IEEE,*
and Hyundong Shin, *Fellow, IEEE*

***Abstract*—Deep reinforcement learning (DRL) has gained significant attention in recent years as a powerful approach for solving complex optimization problems. One of the promising applications of DRL in wireless communication is full-duplex (FD) reconfigurable intelligent surface (RIS)-assisted wireless systems, which have emerged as a potential solution for the next-generation wireless communication networks. FD-RIS-assisted systems can simultaneously transmit and receive data using the same frequency band, which can significantly improve the system capacity and spectral efficiency. This paper provides an overview of the DRL background and its applications in FD-RIS-assisted communication systems. It discusses recent research advances in various scenarios, including resource allocation, sum-rate optimization, and secure communications. Furthermore, it investigates the DRL performance in optimizing large-scale FD-RIS-assisted systems. Major challenges and shortcomings of DRL in FD-RIS-assisted wireless systems are presented and supported through numerical simulations. Based on this discussion, the paper highlights prospective use cases that can bring the FD-RIS-assisted systems into practice.**

***Index Terms*—Reconfigurable intelligent surface, full-duplex, deep reinforcement learning.**

I. Introduction

WITH the rapid development of wireless communication technologies, existing cellular generations are expected to fail the demands of upcoming wireless systems. Therefore, the sixth-generation (6G) wireless systems aims for hybrid technologies that provide unprecedented capabilities, including terabit data rates, microsecond latency, massive device connectivity, and seamless machine learning integration [1]. Existing wireless systems use half-duplex (HD) communications, where transmission and reception occur separately. In contrast, full-duplex (FD) communications allow simultaneous transmission and reception within the same frequency band, thus enabling two-fold spectral efficiency and substantially enhancing the throughput [2]. Despite the promising features of FD communications, their operation properties impose several implementation hurdles at the signal-detection levels. One of the major challenges is self-interference (SI), which occurs when a FD transceiver interferes with its own transmission. The challenge in FD systems is to reduce the SI, such that the receiver can accurately detect the incoming signal. This can be achieved through a combination of propagation, analog, and digital techniques [2]. The other challenge is the co-channel interference (CCI), which is caused by multiple transmitters operating on the same frequency band, leading to overlapping transmissions.

The reconfigurable intelligent surface (RIS) has emerged as a powerful technology for mitigating interference in wireless systems. RIS consists of low-cost passive elements that are independently controlled to manipulate the wavefront of the incoming signal, yielding reduced interference in FD communications. The reflection coefficients can be optimized along with different parameters to maximize key performance metrics such as the sum-rate and energy efficiency, while suppressing the CCI interference. The FD operation has not been beneficial in the case of severe CCI in conventional systems without RIS. However, recent studies have demonstrated that FD-RIS-assisted systems consistently achieve superior performance when compared to HD systems, despite the CCI [3], [4].

Optimizing RIS systems efficiently is a key point to realizing the full potential of FD-RIS-assisted systems. Typically, alternating optimization (AO) techniques are considered, which have proven to achieve promising results. However, such techniques require prior knowledge of the wireless communication environment, along with target-specific mathematical relaxations to achieve near-optimal results. This might not be feasible in dynamic environments. Moreover, these techniques are not scalable. In contrast, deep reinforcement learning (DRL) has evolved as a powerful tool for learning through a trial-and-error approach. Therefore, it can efficiently learn the optimal configuration of the RIS-assisted systems based on the current conditions without the need for mathematical relaxations [5].

Although DRL has been studied within the context of HD wireless systems [5]–[7], its application to FD systems, especially in multi-user environments, remains significantly underexplored. This paper pioneers the investigation into the use of DRL for optimizing FD-RIS-assisted systems, revealing both novel challenges and the potential for substantial performance improvements over traditional algorithms. Specifically, the main contributions are summarized as follows:

- The paper proposes a DRL approach to maximize the sum-rate of multi-user FD-RIS-assisted systems. The results validate the DRL capability in optimizing large state and action spaces, where it achieves near-optimal solutions.
- The paper investigates the problem design challenge in DRL for FD-RIS-assisted systems, where it illustrates the impact of the problem formulation on the DRL performance through simulation results.
- The paper further examines the computational complexity of deep Q-learning (DQL) in the context of FD-RIS-assisted systems, providing insights into the feasibility of implementing DQL-based solutions in such systems.

Besides the main contributions, the paper explores recent research advances in resource allocation, sum-rate optimization, and secure communications within FD-RIS-assisted systems. Furthermore, the major challenges and shortcomings of DRL in FD-RIS-assisted wireless systems are presented and supported by numerical simulations. The paper also highlights prospective use cases that have the potential to bring the FD-RIS-assisted systems into practice. To the best of the authors knowledge, the literature lacks a holistic work of this kind. Towards this end, the next section introduces the powerful DRL tool and provides the distinction from the current optimization tools. Section III details various FD-RIS-assisted systems research advances and use cases. The paper defines open challenges in Section IV. Finally, it recommends potential research directions in Section V and concludes in Section VI.

II. DRL for FD-RIS-aided Wireless Systems

A. Overview

DRL is a sub-field of machine learning that aims at learning the optimal behavior in an environment through a trial-and-error approach. Specifically, the DRL agent performs actions in the dynamic environment and observes the feedback, enabling it to learn decisions that maximize the reward function over time. Unlike supervised or unsupervised learning, DRL provides a feedback signal to the agent that indicates whether its behavior is good or bad, allowing it to adapt without the need for a labeled dataset. This makes it suitable for complex, dynamic, and uncertain environments, such as wireless communication environments, where it may be practically unattainable to specify an accurate solution.

A DRL problem can be identified using the following key elements: *state*, *action*, *reward*, *policy*, and *value function*. The state represents the current configuration of the environment. The action represents the decisions that the agent aims to optimize. The reward defines the goal of a DRL problem, where the environment sends a scalar value (i.e., the reward) as feedback to evaluate the action taken by the agent. The policy maps the current state of the environment to actions. Depending on the problem, the policy can be expressed through a lookup table, or it might require complex computations in other instances. Specifically, the policy can be represented by a neural network (NN) that takes the environment states as inputs and approximates the actions. The value function identifies the action evaluations on long-term goals.

The learning process of the DRL agent is formulated as episodes and steps. An episode refers to a complete series of interactions between the agent and the environment, from the initial state to the terminal state. In each episode, the individual actions taken by the agent represent the steps in response to the current state. In each step, the agent chooses an action to perform based on the environmental state and policy. It observes the reward, and transitions to the next state. This continuous interaction enables the agent to learn about the environment properties and determine the optimal actions in real-time. In wireless communications, a specific terminal state may not naturally exist as in gaming tasks. Therefore, the number of steps and episodes can be set as a fixed number, allowing the learning process to continue until convergence is achieved.

B. AO vs. DRL

The optimization of FD-RIS-assisted systems has been extensively investigated in the literature using AO approaches. However, such approaches pose several practical implementation challenges. Particularly, AO techniques demand well-established prior mathematical models that vary depending on the system model and objective function, which can be difficult to attain in large-scale systems. It further restricts the objective formulation to convex problems, requiring several relaxation steps that can result in far sub-optimal solutions. Conversely, DRL can handle non-linear, non-convex, and high-dimensional problems without the need for prior mathematical relaxations.

Furthermore, DRL can learn to adapt to dynamic environments. The RIS-assisted wireless propagation environment can vary due to different factors, such as the user equipments (UEs) mobility, change in the number of UEs, presence of new obstacles, weather fluctuations, and randomness of channel state information (CSI). By updating the policy and adjusting the RIS configuration according to the value function, the agent becomes robust against system changes. Moreover, DRL searches for the optimal strategy without the need for prior knowledge of the propagation environment. In contrast, conventional AO approaches may not be able to seamlessly find the best decision in a probabilistic setting. It relies on prior knowledge to achieve optimal performance and is usually obtained through extensive measurements or mathematical modeling. This poses several limitations to practical deployments, including the need for accurate models, which can be difficult to obtain in dynamic or complex propagation environments.

Beyond the learning capabilities, DRL can optimize multiple objectives efficiently in different systems, including the sum-rate, system power, and secrecy rate. In particular, DRL only needs to design an appropriate reward function that simultaneously optimizes multiple targets. In contrast, AO techniques need to divide the problem into sub-problems and iterate through the objectives until convergence. Given that each sub-problem requires relaxations and assumptions about the propagation environment, the running time and reliability of the system cannot be maintained. Additionally, DRL is particularly well-suited for FD systems because of its ability to efficiently optimize the RIS configuration while dynamically managing CCI, which can severely degrade the performance if not properly managed.

III. Research Advances

Due to the above features, DRL is envisioned to be powerful for complex wireless systems, such as the optimization problems of FD-RIS-assisted networks. Figure 1 illustrates some use cases of FD-RIS-assisted systems, which will be discussed in the context of various optimization objectives later in this section.

A. Single-user Sum-rate Optimization

Recently, DRL has emerged as a promising tool to optimize the sum-rate of FD-RIS-assisted systems [8]–[11]. In what follows, the literature works are discussed based on the decision parameters categorization: continuous and discrete.

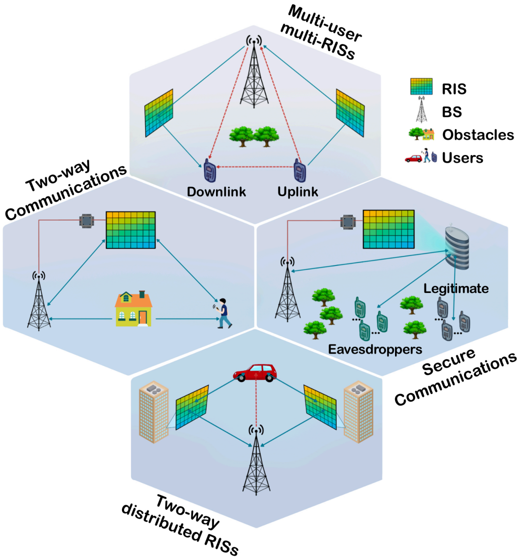
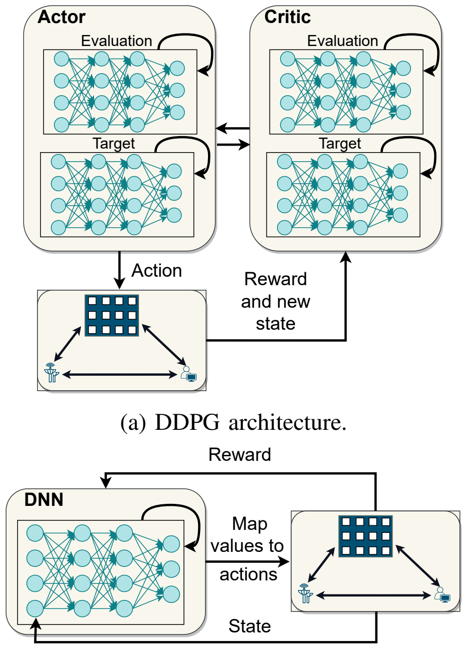


Fig. 1: FD-RIS-assisted communication system use cases.

1) Continuous Action Space: In [8], the authors investigated the RIS-assisted multiple-input single-output (MISO) single-user system, as shown in Fig. 1. Two operating modes are considered: HD and FD. The goal is to optimize the continuous RIS phase shifts and transmit beamformers to maximize the system-rate. The deep deterministic policy gradient (DDPG) is leveraged for optimizing the RIS phase shifts as it is well-suited for continuous action spaces. The DDPG consists of two main networks: actor and critic. The former learns a deterministic policy, which maps the state directly to an action value. The latter learns to approximate the action-value function, estimating the expected total reward starting from a given state and following a given policy. Both networks are implemented using deep NN (DNN).

Furthermore, the DDPG uses experience replay and target networks to best learn the optimal policy through interaction with the wireless environment. The experience replay stores transitions from previous experiences, and the target network is used to stabilize the training process. The experience, represented by the state, action, reward, and next state, is stored in the experience replay. The agent randomly chooses a batch of transitions from the experience replay to update the networks. The process is repeated for a number of episodes and steps until convergence. In [8], the state space included the RIS continuous phase shifts and reward function, the action included the RIS phase shifts, and the reward was defined by the problem goal (i.e., the sum-rate function). To this end, the DDPG algorithm architecture is shown in Fig. 2a. The work in [8] showed that the DRL algorithm remarkably enhances the system-rate, in both HD and FD modes, compared to the non-optimized case using a unified DRL parameter setting. Furthermore, the work in [9] investigated the single and distributed RIS deployment schemes in an FD-RIS-assisted MISO system. The target was to maximize the sum-rate by optimizing the beamformers and continuous RIS phase shifts to investigate the preference of deploying a single or distributed RIS using three practical scenarios. The paper showed that the DDPG can be efficiently deployed to optimize decision parameters based on different system settings. The work in [12] investigated the sum-rate maximization problem of an FD-RIS-assisted vehicular communication system. The paper demonstrated that DRL can mitigate residual SI and CCI issues in FD systems, thereby facilitating more reliable FD-RIS-assisted networks in vehicular settings.



(a) DDPG architecture.

(b) DQL architecture.

Fig. 2: DRL algorithms.

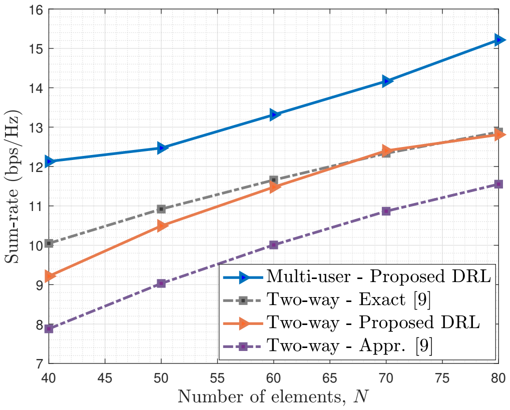


Fig. 3: DDPG performance in multi-user FD-RIS-assisted system. UEs: 8, antennas at the BS: 8, experience replay size: $100,000$, batch size: 16, discount factor: 0.99, learning rate: 0.001, soft update: 0.001, episodes: 500, and steps: 8000.

2) Discrete Action Space: Beyond the continuous system setting, few works considered a practical (i.e., discrete) RIS phase shift model to combat the hardware limitations of the continuous model. DQL is one of the powerful DRL frameworks that can be used efficiently to optimize discrete action spaces. The work in [11] considered a distributed RIS-assisted two-way single-user communication system as illustrated in Fig. 1, where the discrete phase shifts restriction at the RIS is investigated. The goal was to maximize the sum-rate by optimizing the beamformers and discrete RIS phase shifts. As there exist closed-form mathematical solutions for the beamformers, the paper leveraged DQL for the phase shift optimization. DQL has the ability to handle high-dimensional state spaces using DNN and to learn optimal policies without comprehending the system dynamics. It aims at approximating the Q-function, which estimates the value of each action in a given state. Typically, the agent chooses the action that has the maximum Q-value. However, the agent also chooses random actions to prevent saturating at a local minimum. The structure of the DQL algorithm is shown in Fig. 2b. The paper proved the practicality of DQL in such use case, where it almost achieved the continuous phase shift model performance with only a resolution of 6 bits.

Having discussed different algorithms to maximize the same objective, one can observe that choosing an appropriate DRL algorithm for FD RIS-assisted systems is important, which depends critically on the characteristics and requirements of the problem. Specifically, Q-learning is well-suited for environments with a discrete action space. However, Q-learning can struggle with high-dimensional state and/or action spaces because it requires a lookup table to find the optimal solution. DQL extends Q-learning by using a DNN to approximate the Q-value function, thus enabling the handling of large-scale environments. The DDPG is designed to operate over continuous action spaces and is ideal for scenarios where actions require continuous tuning, such as power levels or beamformers in FD systems.

B. Proposed DRL Algorithm for Multi-user Sum-rate Optimization

In multi-user wireless systems, the action and state spaces are typically high-dimensional, making it difficult to achieve optimal solutions using AO techniques. However, the DDPG can handle large state and action spaces using DNN, experience replay, and target networks. The actor-network maps the current state to a continuous action space, allowing it to handle different decision parameters, including continuous RIS phase shifts, beamformers, and power in FD-RIS-assisted communication systems. All of the literature work only included one decision parameter in the action space (i.e., the RIS phase shifts). Since there are no mathematical closed-form solutions for optimizing the beamformers in multi-user communication systems, powerful algorithms that achieve near-optimal performance should be developed.

To validate the performance of the DDPG in FD-RIS-assisted systems, we investigate the sum-rate optimization of a multi-user FD-RIS-assisted system. We jointly optimize the continuous RIS phase shifts and beamformers to maximize the system sum-rate. The optimization problem is challenging due to the multi-user interference, in addition to the extended state and action scales. Therefore, the DRL formulation should be designed accurately, along with several performance improvement techniques. The construction of the action, state, and reward is as follows: The action space contains the beamforming matrix and RIS phase shifts. The state is determined by the transmission power, received power of UEs, action from the previous step, and CSI. We consider the reward function to be the difference between the previous and current step sum-rate to encourage the agent to improve its performance by receiving positive rewards for a rate increase and negative rewards otherwise. The actor and critic networks have an input layer, a hidden layer, and an output layer. The actor-network hidden layer is modeled as long short-term memory, while the critic-network hidden layer is modeled as a feedforward network. Input sizes correspond to state space and concatenated action/state spaces, respectively. The actor-network output matches the action space, while the critic-network has one neuron. Both networks hidden layer consist of 512 neurons with tanh activation function. All complex-valued parameters are separated as real and imaginary parts and fed into the DNN independently. Moreover, to enhance the efficiency of the function approximation, the state is whitened prior to passing it to both the critic and actor networks. Batch normalization is also applied at the hidden layers to stabilize the system output.

Figure 3 shows the DDPG sum-rate performance compared to the number of reflecting elements. The algorithm proposed in [9] is included as a benchmark for comparison in the single-user FD setting. In [9], the DDPG is adopted to optimize the RIS phase shifts, and two formulations of closed-form derivations are used to optimize the beamformers: exact and approximate solutions. As can be observed in the two-way communication setting, the proposed joint optimization achieves near-optimal performance efficiently, and it outperforms the approximate closed-form solutions. It further shows that the gap between the exact closed-form solution and the DRL joint optimization decreases as the number of RIS elements increases. It is worth noting that the proposed DDPG approximates the beamformers and RIS phase shifts jointly without the need to iterate through two sub-problems (i.e., phase shift optimization using DRL and beamformers optimization using closed-form solutions). The simulation result proves that the DDPG can be efficiently exploited to optimize various scenarios, given the accurate DRL design and enhancement techniques.

C. System Resource Optimization

Besides the sum-rate maximization problems, the DRL approaches can be applied to various objectives and setups. Particularly, the work in [13] considered minimizing the FD-RIS-assisted system resources by optimizing the discrete RIS phase shifts and their states (either ON or OFF). Assuming that all RIS elements are constantly ON can lead to significant resource wastage, as all practical applications function at a specific target rate. This assumption in practice results in higher power usage, channel estimation efforts, and resources for optimizing phase shifts. Therefore, the authors considered the deployment of the DQL to minimize the system resources by optimizing two discrete action spaces, the discrete RIS phase shifts and their states.

Consequently, the DRL formulation would be approached differently than the above formulations. The reward function cannot be simply based on the objective function. From the agent perspective, the goal is to select actions that maximize the cumulative reward, which may result in activating most elements and surpassing the target sum-rate. Therefore, to ensure the optimal number of active elements that achieves the system target rate are selected, the reward function should incorporate the target rate constraint. Hence, the reward function is designed using a combination of punishment and encouragement strategies. Negative rewards are applied if the agent does not meet the target rate constraint, while positive rewards are assigned when it meets the target rate constraint and enhances the rate based on prior feedback. It is shown in the paper that the DQL algorithm effectively satisfies the target rate and deactivates a considerable number of RIS elements. This proves that the DQL is applicable to different scenarios, each with its own target rate requirements.

D. Secrecy Rate Optimization

In addition to the above formulations, RISs can decrease the data rate at eavesdroppers while augmenting the data rate at legitimate receivers, thus improving the overall secrecy rate. Incorporating the FD technology in the RIS-assisted communication provides the possibility of enhancing the physical layer security, where it is challenging for the eavesdropper to decode the signal. Therefore, optimizing secrecy rates and RIS design in an FD communication setting is of paramount importance. The work in [14] proposed a DRL-based approach for RIS-aided multi-user FD secure systems under hardware impairments, as shown in Fig. 1. The paper considered a practical scenario where the RIS is equipped with continuous phase shifts that suffer from quantization errors and mutual coupling between the phase shifts. The proposed DRL algorithm takes these hardware impairments into account while jointly optimizing the RIS phase shifts and beamformers. The results demonstrated that the developed DRL approach can effectively enhance the system sum secrecy rate and outperforms existing baseline schemes.

IV. Challenges

A. Problem Design

DRL has exhibited outstanding performance in various RIS-assisted wireless applications. Yet, developing an effective DRL algorithm is a challenging task that involves designing an accurate problem formulation that suits the capabilities of the DRL algorithms. The complexity of the communication system and the limitations of the DRL algorithm, such as the exploration-exploitation trade-off and the curse of dimensionality, can impose several implementation hurdles. Therefore, developing an optimal problem design that can leverage the strengths of DRL algorithms and overcome their limitations is vital to achieving the desired performance in FD-RIS-assisted systems. Fig. 4 illustrates the importance of a proper reward design in achieving the target rate. The simulation result considers the system model and problem formulation presented in Sec. III-C, in which the target rate is set to 10 bps/Hz. Two reward designs are investigated in addition to the design proposed in [13], denoted by Case 1. The agent is considered to have satisfied the target rate constraint if the sum-rate exceeds the target rate but remains less than the upper bound, which is defined as target rate + 1. Case 2 assumes a smaller target rate upper bound, of 0.5 instead of 1. Case 3 excludes the negative penalty from [13], leaving the agent to learn from the two extremes. Although the reward designs are not technically wrong, the agent fails to satisfy the objective of the problem with improper reward designs. This not only impacts the overall performance but also significantly hinders the convergence of the DRL algorithm, delaying or even preventing it from reaching an optimal solution. Consequently, the reward design should be carefully tailored to the constraints and operational dynamics of FD-RIS-assisted systems, considering factors such as interference management, power efficiency, and real-time adaptation capabilities.

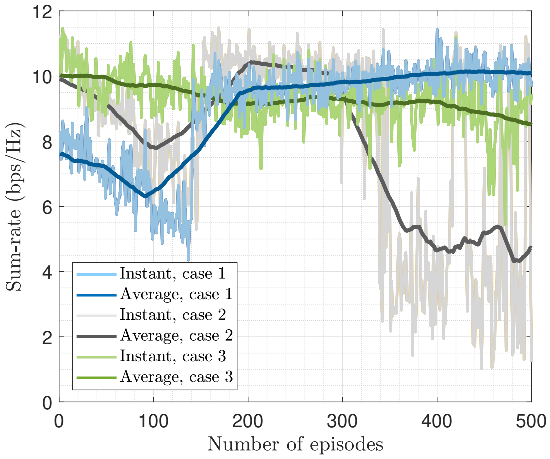


Fig. 4: The impact of the reward design on the performance.

B. Computational Complexity

DRL has emerged in different fields by tackling complex and large-scale problems efficiently. A notable example is the AlphaGo project, developed by DeepMind Technologies, which used DNNs and Monte Carlo tree search algorithms to master the game of Go [15]. In 2016, AlphaGo competed against the world champion and won four games, marking a milestone in the development of artificial intelligence. However, a single game of AlphaGo required 1,920 central processing units and 280 graphics processing units with an estimated cost of $35,000,000. Given the large parameter space, and the lack of solid complexity analysis in existing literature, the question of whether we can fulfill the computational requirements of DRL deployments in large-scale RIS-assisted systems remains an open challenge that needs to be addressed. However, despite the extensive memory requirements, DRL algorithms are able to perform decision-making with minimal latency during the testing phase, similar to traditional algorithms, ensuring that real-time responsiveness is maintained even in complex environments.

Figure 5 shows the impact of increasing the number of RIS elements on the complexity of the DQL algorithm. The NN parameters and simulation settings are based on [11]. As observed, the time complexity, which is based on the total number of episodes and steps, is constant even for larger numbers of reflecting elements. The number of additions and multiplications increases slightly as the number of reflecting elements increases. In contrast, the space complexity of DQL (i.e., represented by the number of parameters) increases rapidly as the number of reflecting elements increases. The

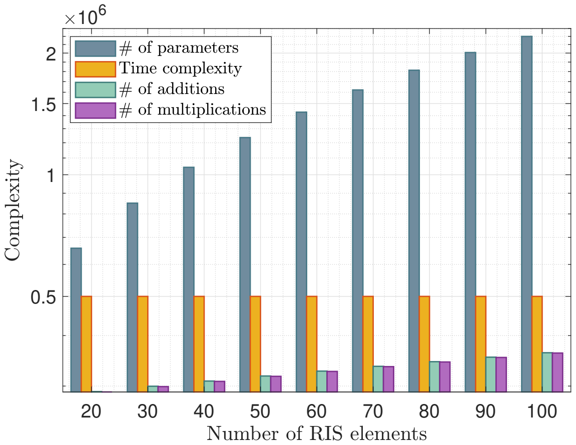


Fig. 5: The impact of increasing the number of RIS elements on DQL complexity.

network evaluates the Q-value for each possible action, which is computationally expensive and can increase the memory requirements of the algorithm.

V. Prospective Research Directions

Having detailed the research trends and challenges of DRL in FD-RIS-assisted communications, the paper now discusses a few prospective use cases that are not yet investigated in the literature.

A. Large-scale FD-RIS-assisted Communication Systems

Although some of the practical FD settings are investigated in the literature, there are several important problems that remain an open challenge. Most of the investigated problems considered a single-user simulation setting, which enabled the action and state spaces (either continuous or discrete) to be consistent. However, when considering mixed decision parameters, effective DRL algorithms and techniques should be developed to achieve optimal performance. An example of a mixed optimization problem is the joint optimization of discrete variables (e.g., discrete RIS phase shifts and states) and continuous variables (e.g., power and beamformers). Several techniques have emerged in the field of DRL to tackle such problems. For example, hybrid DDPG can be considered, which combines both the DDPG and DQL algorithms to respectively handle continuous and discrete actions in a unified framework. The continuous actions are learned using the DDPG algorithm, which uses an NN to represent the policy. The discrete actions are learned using DQL, which employs a separate NN to represent the Q-function.

Furthermore, integrating large language models (LLMs) with DRL in large-scale systems offers a robust approach to managing the complex relations and numerous network elements. Specifically, LLMs can be used to dynamically adjust the reward functions used in DRL, based on the current network state or desired target, such as minimizing interference or maximizing throughput. They can also help in translating complex network optimization tasks into DRL objectives by interpreting the technical specifications and constraints described in network policy. By blending the power of LLMs with the adaptive learning capabilities of DRL, networks can achieve higher levels of optimization and operational efficiency, which is crucial for handling the scale and complexity of future wireless networks.

B. Meta-reinforcement Learning (MRL)

Despite remarkable developments, traditional RL methods cannot quickly adapt to new tasks using prior knowledge. In MRL, agents are designed to leverage prior experience on similar tasks. MRL focuses on learning to learn. There are key components that differentiate RL from MRL, which are summarized as follows: MRL needs to use a model with memory to acquire and store knowledge about the current task from the immediate environment, which would help to update its hidden state. Furthermore, a MRL algorithm defines how the model weights are updated based on what it learned. The main objective of the algorithm is to help optimize the model to solve an unseen task in the minimum amount of time, applying the prior knowledge. The MRL system includes the last reward and action in the policy observation along with the current state. The purpose of this is to feed and keep track of the history of all tasks and observations so that the model can internally update the dynamics between the states, actions, and rewards based on the current configuration. This approach is particularly useful when data availability is limited, as it allows the agent to efficiently use the available data by building upon prior knowledge and adapting quickly to new tasks.

C. Quantum DRL (Q-DRL)

Q-DRL represents a cutting-edge fusion of quantum mechanics theories and DRL approaches. It employs quantum mechanics principles, such as superposition and entanglement, to tackle complex decision-making problems efficiently in large-scale probabilistic environments. In FD-RIS systems, efficient resource allocation, beamforming, and interference management are essential for enhancing the performance. The Q-DRL leverages the quantum properties of qubits to explore the vast solution space more efficiently. Besides the scalability advantage, the Q-DRL can tackle the limitations of the classical DRL algorithms. Specifically, classical DRL algorithms often struggle to strike the balance between exploitation and exploration. Q-DRL benefits from its ability to explore multiple actions concurrently, which leads to a stable environment exploration. Similarly, the experience replay in classical DRL involves sampling past transitions to train the agent. Selecting the training samples plays a key role in the agent learning process. Q-DRL can exploit quantum parallelism to examine multiple experiences simultaneously, thus speeding up the training process effectively, which is particularly useful for FD-RIS-assisted systems.

D. DRL for FD-RIS-Assisted Integrated Sensing and Communications (ISAC)

ISAC has emerged as an important function for the 6G wireless networks. Many practical applications not only require high-quality communication but also need simultaneous localization with high precision. Therefore, the concept of ISAC was invented to best use the system resources (e.g., spectrum, energy, and hardware) through shared sensing and communication (SAC) functions. Incorporating RIS technologies into ISAC systems provides more degrees of freedom and optimized performance through RIS phase shift optimization. Therefore, by jointly designing the RIS phase shifts, beamformers, and power using DRL approaches, one may leverage the RIS capabilities to facilitate efficient ISAC transmission. The use of RIS in FD-based ISAC systems not only enhances SAC performance but also plays a crucial role in managing the potential interference inherent in FD operations, such as SAC interference and CCI between simultaneous downlink and uplink transmissions.

In addition to the benefits of RIS in ISAC systems, the application of DRL extends beyond conventional RIS types to other variants such as simultaneous transmit and reflect RIS (STAR-RIS), which offer enhanced capabilities. STAR-RIS, not only tunes phase shifts but also divides signals into separate paths, aligning perfectly with the dual demands of ISAC for high-precision localization and robust communication. The integration of DRL with STAR-RIS and other RIS configurations such as hybrid and active RIS, each providing unique advantages in signal control and manipulation, could further revolutionize FD operation modes in ISAC systems. DRL capability to dynamically adapt and optimize the allocation of resources in real-time becomes crucial in exploiting these RIS functionalities.

E. Non-terrestrial Networks (NTNs)

The vision for 6G marks an evolution from wireless standards, transitioning from ground-based coverage to an integrated approach that includes aerial technologies. Utilizing NTNs is essential to achieve global connectivity, and introduces promising prospects for FD communications.

In aerial networks, the difference in power levels between transmitted and received signals tend to be smaller than that of ground-based systems due to shorter transmission distances. This reduced power imbalance at low altitudes enables combating the SI issue of FD communications, thus enhancing the network performance. Building on the advancements of NTNs and FD communications at lower altitudes, the integration of RIS can enable more precise control of signal paths to enhance signal quality. Furthermore, RIS is expected to be one of the most cost-efficient solutions to address NTNs practical issues including overload power consumption and high probability of blockage. However, the movement of aerial terminals and the dynamic nature of NTNs pose considerable implementation challenges. Hence, integrating space segments with terrestrial networks would require thorough planning. DRL presents a promising solution to address the inherent challenges posed by the mobility of aerial platforms, such as time-varying interference patterns. It can adaptively learn and optimize network parameters in real-time, making it ideal for managing the dynamic environmental conditions encountered in aerial networks.

VI. Conclusion

This paper provided a holistic overview of the application of DRL to FD-RIS-assisted systems. It highlighted the main positive aspects of DRL as compared to the traditional optimization methods. Furthermore, it defined multiple research advances that are critical for exploiting the DRL capabilities in optimizing FD-RIS-assisted systems. It further investigated the scalability of DRL algorithms in optimizing multi-user FD-RIS-assisted systems. Some of the key challenges that face the implementation of DRL were discussed and supported through numerical simulations. Finally, the paper advocated for potential research directions that should be investigated to realize the full potential of DRL in FD-RIS-assisted wireless communication systems in the near future.

References

[1] C.-X. Wang *et al.*, “On the road to 6G: Visions, requirements, key technologies, and testbeds,” *IEEE Commun. Surv.*, vol. 25, no. 2, pp. 905–974, Feb. 2023.

[2] D. Bharadia *et al.*, “Full duplex radios,” in *Proc. ACM SIGCOMM*, Aug. 2013, p. 375–386.

[3] Y. Cai *et al.*, “Intelligent reflecting surface aided full-duplex communication: Passive beamforming and deployment design,” *IEEE Trans. Wirel. Commun.*, vol. 21, no. 1, pp. 383–397, Jan. 2022.

[4] R. Sultan and A. Shamseldeen, “Uplink–downlink cochannel interference cancellation in RIS-aided full-duplex networks,” *IEEE Syst. J.*, pp. 1–4, Apr. 2024.

[5] R. Hashemi *et al.*, “Deep reinforcement learning for practical phase-shift optimization in RIS-aided MISO URLLC systems,” *IEEE Internet Things J.*, vol. 10, no. 10, pp. 8931–8943, May 2023.

[6] K. Cao and Q. Tang, “Energy efficiency maximization for ris-assisted miso symbiotic radio systems based on deep reinforcement learning,” *IEEE Commun. Lett.*, vol. 28, no. 1, pp. 88–92, Apr. 2024.

[7] Q. Liu *et al.*, “DRL-based secrecy rate optimization for RIS-assisted secure ISAC systems,” *IEEE Trans. Veh. Technol.*, vol. 72, no. 12, pp. 16 871–16 875, 2023.

[8] A. Faisal *et al.*, “Deep reinforcement learning for optimizing RIS-assisted HD-FD wireless systems,” *IEEE Commun. Lett.*, vol. 25, no. 12, pp. 3893–3897, Dec. 2021.

[9] ——, “Deep reinforcement learning for RIS-assisted FD systems: Single or distributed RIS?” *IEEE Commun. Lett.*, vol. 26, no. 7, pp. 1563–1567, Apr. 2022.

[10] Z. Peng *et al.*, “Multiuser full-duplex two-way communications via intelligent reflecting surface,” *IEEE Trans. Signal Process.*, vol. 69, pp. 837–851, Jan. 2021.

[11] A. Faisal *et al.*, “Distributed RIS-assisted FD systems with discrete phase shifts: A reinforcement learning approach,” in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2022, pp. 5862–5867.

[12] P. Saikia *et al.*, “Proximal policy optimization for RIS-assisted full duplex 6G-V2X communications,” *IEEE Trans. Intell. Veh.*, pp. 1–16, May 2023, DOI: 10.1109/TIV.2023.3275632.

[13] A. Faisal *et al.*, “On discrete phase shifts optimization of RIS-aided FD systems: Are all RIS elements needed?” in *Proc. 2023 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2023, pp. 1–6.

[14] Z. Peng *et al.*, “Deep reinforcement learning for RIS-aided multiuser full-duplex secure communications with hardware impairments,” *IEEE Internet Things J.*, vol. 9, no. 21, pp. 21 121–21 135, Nov. 2022.

[15] D. Silver *et al.*, “Mastering the game of go with deep neural networks and tree search,” *Nature*, vol. 529, pp. 484–503, Jan. 2016.