When Optimization Meets Machine Learning: The Case of IRS-Assisted Wireless Networks

Shimin Gong, Jiaye Lin, Beichen Ding, Dusit Niyato, Dong In Kim, and Mohsen Guizani

Abstract

Performance optimization of wireless networks is typically complicated because of high computational complexity and dynamic channel conditions. Considering a specific case, the recent introduction of intelligent reflecting surface (IRS) can reshape the wireless channels by controlling the scattering elements' phase shifts, namely, passive beamforming. However, due to the large size of scattering elements, the IRS's beamforming optimization becomes intractable. In this article, we focus on machine learning (ML) approaches for complex optimization problems in wireless networks. ML approaches can provide flexibility and robustness against uncertain and dynamic systems. However, practical challenges still remain due to slow convergence in offline training or online learning. This motivated us to design a novel optimization-driven ML framework that exploits the efficiency of model-based optimization and the robustness of model-free ML approaches. Splitting the control variables into two parts allows one part to be updated by the outer loop ML approach while the other part is solved by the inner loop optimization. The case study in IRS-assisted wireless networks confirms that the optimization-driven ML framework can improve learning efficiency and the reward performance significantly compared to conventional model-free ML approaches.

Introduction

Intelligent reflecting surface (IRS) was recently introduced in wireless communications to configure the signal propagation environment in favor of information transmission [1]. By a joint phase control of the IRS's scattering elements, namely, passive beamforming, the signal reflections can be tuned to create desirable channel conditions. The radio environment can be optimized to improve network performance (e.g., channel capacity and transmit power savings). The IRS's passive beamforming optimization is challenged by the computational complexity with a large number of scattering elements. Moreover, it relies on the exact channel information, which becomes more difficult to estimate for passive scattering elements with limited information processing capability. As such, it becomes difficult for beamforming optimization in IRS-assisted wireless networks, which generally requires model simplification, approximation, and sophisticated algorithm design. Besides model-based optimization solutions, we observe an upsurge of model-free machine learning (ML) approaches, which is also proposed for performance optimization problems in IRS-assisted wireless networks. However, the integration of model-based optimization and model-free ML approaches is seldom studied. The optimization solutions typically rely on the exploration of the structural information. For example, the deterministic and functional connections of control variables due to their physical dependence. We envision that such structural information can also be exploited to improve the learning performance of conventional ML approaches.

Motivated by the above observations, our contributions to this article are twofold. First, we propose a novel optimization-driven ML framework to solve complex problems in wireless systems. We utilize optimization solutions to drive the learning process toward a better target in a more efficient way. It is expected that we improve the convergence speed and get closer to the optimal solution compared to the conventional model-free ML approach. Second, to verify the proposed learning framework, we present a case study for joint active and passive beamforming optimization in IRS-assisted wireless networks. The remainder of the paper is organized as follows. In the following section, we provide an overview of IRS-assisted wireless networks, emphasizing the performance gains and challenging the issues. Then we review different applications of ML approaches in IRS-assisted wireless networks. Comparing the optimization methods, ML approaches provide enhanced flexibility and robustness against uncertain channel information and inexact system modeling. We also reveal some practical challenges to deploy ML approaches, mainly due to the requirement of offline training or slow convergence in online learning. We design an optimization-driven deep reinforcement learning (DRL) framework to exploit the efficiency of model-based optimization methods and the robustness of model-free DRL approaches. The basic idea is to split the control variables of a complex problem into two parts. One part can be learned in the outer-loop DRL approach, while the other part can be optimized

Shimin Gong, Jiaye Lin, and Beichen Ding (corresponding author) are with the Shenzhen Campus of Sun Yat-sen University, China; Dusit Niyato is with Nanyang Technological University, Singapore; Dong In Kim is with Sungkyunkwan University, South Korea; Mohsen Guizani is with Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), UAE.

Digital Object Identifier:
10.1109/MNET.211.2100386

0890-8044/22/$25.00 © 2022 IEEE

by solving an approximate problem efficiently. As such, we can reduce the search space in the outer-loop DRL approach and expect an improvement in the learning efficiency. Following that, the case study reveals that the new approach can achieve significant improvement by using a smaller number of learning episodes, compared to the model-free DRL approach. Finally, we discuss some future research direction.

An Overview of IRS-assisted Wireless Systems

The reconfiguration of the IRS relies on tunable chips embedded in its structure. Each tunable chip is controllable to adapt the phase shift of each scattering element. The performance improvement of IRS-assisted wireless networks can be achieved by using IRS as either a signal reflector, a signal transmitter, or a signal receiver. In the following, we discuss different uses of IRS and review design challenges in IRS-assisted wireless networks.

Signal Reflector: The IRS can be used to reflect the incident radio frequency (RF) signals and create an additional link to the receiver. The RF signals in the direct link can be combined with its reflections at the receiver, or combined destructively to suppress information leakage to unintended receivers. The enhanced channel condition implies significant power saving at the RF transmitter. The power scaling law in [2] shows that the transmit power of an IRS-assisted base station can be scaled down in the order of $1/N^2$ without compromising the receiver's performance, where N denotes the size of the IRS. The IRS's reflections can also be used to suppress information leakage to illegitimate users. It can be more effective to enhance secrecy rate and energy efficiency by deploying a large-size IRS instead of increasing the size of antenna array at the RF transmitter [3].

Signal Transmitter: By controlling the phase shifts of the IRS dynamically, the signal reflections can exhibit different radiation patterns and be used to carry useful information. This can be viewed as a form of backscatter communications underlying the RF communications [4]. By using a large number of scattering elements, the IRS-based backscatter communications can generate more exotic reflection patterns for information communications, leading to a higher data rate and a larger transmission range.

Signal Receiver: The scattering elements of the IRS can also be used as individual receivers for multiuser simultaneous data transmission. The IRS can achieve an impressive capacity gain by suppressing the interference among different transmitters. The large IRS can be divided into smaller units, processing individually the received signals from different transmitters. The IRS can also be used as an array of sensors to estimate the position of RF transmitters, based on the received signal strengths at the IRS's scattering elements.

A joint active and passive beamforming optimization is often required to achieve the performance gains in terms of channel capacity, transmit power savings, or secrecy enhancement. The joint beamforming optimization is typically solved by alternating optimization that decomposes the IRS's phase control and the active beamforming into two sub-problems. In each sub-problem, semidefinite relaxation is usually required to optimize the beamforming matrix by solving a semidefinite program. However, the optimization-based methods suffer from the following practical difficulties:

- Computational complexity: A larger size of the scattering elements allows more flexible reconfigurability. It also incurs a higher computational complexity when passive beamforming optimization is coupled with the transmitter's active beamforming. This often requires more sophisticated algorithm designs with convex approximations and decompositions.
- Dynamic channel conditions: The channel estimation is required within each coherent time interval for efficient and precise beamforming optimization. With a large IRS, more training overhead is required and the error estimates are inevitable. Hence, the optimization methods become costly and unstable due to dynamic and uncertain channel information.
- Imprecise modeling: The beamforming optimization is typically built on a simplified system model (e.g., static channel conditions and continuous phase control). An efficient algorithm design requires convex approximations or heuristic decompositions. As such, the optimization solution only provides an approximation to the original design problem.

The above challenges motivate the use of model-free ML approaches to solve the joint beamforming optimization problem. In the sequel, we first review the applications of ML approaches in IRS-assisted wireless systems (Table 1) and then analyze the current limitations, which motivate us to design a novel optimization-driven ML framework below.

The reconfiguration of the IRS relies on tunable chips embedded in its structure. Each tunable chip is controllable to adapt the phase shift of each scattering element. The performance improvement of IRS-assisted wireless networks can be achieved by using IRS as either a signal reflector, a signal transmitter, or a signal receiver.

Machine Learning for IRS-Assisted Wireless Systems

ML approaches include supervised and unsupervised learning, depending on the availability of labeled samples in the training data. The learning performance can be improved by leveraging deep neural networks (DNNs) to extract hidden features of the training data. Reinforcement learning (RL) makes decisions by continuously interacting with the uncertain environment, which can be modeled by a Markov decision process with a properly defined system state, action set, and the reward function. DRL improves the RL's learning performance by using DNNs to approximate different components of the RL framework. In particular, deep Q-network (DQN) uses DNNs to approximate the value function. Deep deterministic policy gradient (DDPG) uses two sets of DNNs, namely, the critic and actor networks, to estimate the value function and the policy function, respectively. In the sequel, we briefly review the applications of ML approaches in IRS-assisted wireless systems.

REF	System	ML approach	Objective	Conclusions
[5]	MIMO	CNN	Channel estimation	More robust performance
[6]	MIMO	DNN	Channel estimation	Achieve a considerable error performance with a small number of active elements
[7]	OFDM	DNN	Data rate	Achieve the upper bound with perfect CSI
[8]	MISO	GNN	Sum rate	Provide generalizability and improve sum-rate performance with fewer pilots
[9]	MISO	DNN	Received SNR	Comparable with the SDR-based optimization
[10]	MISO	DQN	Energy efficiency	Energy efficiency increased by 77.3 percent
[11]	MISO	DDPG	Received SNR	Close-to-optimal SNR with low time overhead
[12]	MISO	DDPG	Sum rate	Comparable with optimization methods
[13]	MISO	DQN	Secrecy rate	Improved secrecy rate and quality of service

TABLE 1. ML applications in IRS-assisted wireless networks.

Supervised Learning for IRS-Assisted Channel Estimation: The channel estimation can be performed in a training period by sending a known pilot and then estimating the channel information based on the channel response at the receiver. It is typically performed at one end point of the communication process — for example, the access point with higher computational capability. The input pilot and the expected channel response can be viewed as the labeled data for supervised learning. The IRS's channel estimation generally assumes one active scattering element in each training period, while all the other elements are inactive. Hence, a large-size IRS will generate a huge data set during channel training, which can be handled by data-driven ML approaches. For example, a convolutional neural network (CNN) is employed in [5] to estimate both direct and cascaded channels for an IRS-assisted system based on simulated input signals and the expected output channel vectors. The well-trained CNN is then used to predict the real-time channel conditions. To reduce the training overhead, the authors in [6] designed a hybrid IRS architecture and proposed DNN for channel estimation in a millimeter-wave system.

Supervised/Unsupervised Learning for Beamforming Optimization: The beamforming optimization in an IRS-assisted system depends on the wireless environment between the RF transceivers and the IRS. DNNs can be trained offline to recall high-dimensional mapping from the environmental features (e.g., channel response and the receiver's location) to the optimal beamforming strategy that maximizes the received signal strength [7]. The well-trained DNN is then used for online prediction of the IRS's optimal phase vector, given the receiver's location. The graph neural network is employed in [8] to provide better scalability and generalization for beamforming optimization when the network conditions change. Unsupervised learning can also be used for beamforming optimization to maximize the signal-to-noise ratio (SNR) at the receiver by properly designing the loss function [9].

DRL for Beamforming Optimization: The RL/DRL approaches allow online decision-making by interacting with the environment. Considering a discrete action space, DQN can be used to update the IRS's phase shifts based on the observed channel conditions and the receiver's feedback [10]. The continuous phase vector can be directly optimized by DDPG to maximize the received SNR [11] or the sum rate [12]. These works reveal that DDPG can achieve comparable performance to that of the conventional optimization-based algorithms. DDPG can also be used to enhance physical layer security of IRS-assisted systems by beamforming optimization to minimize information leakage to eavesdroppers [13].

Compared to optimization-based methods, the ML approaches in IRS-assisted systems demonstrate more flexibility and robustness against uncertain information, imprecise modeling, and dynamic environment. However, their practical implementations are still challenging, mainly due to the requirement of offline training or slow convergence in online learning. In particular, the DNN trainings in [5–9] rely on a sufficiently large set of training data. The training data is typically generated from simplified models, which may introduce systematic bias for online prediction. Though RL/DRL methods learn to make decisions from scratch [10–13], they are subject to slow convergence by interacting with the environment. The online exploration becomes inefficient with large action and state spaces.

Optimization-Driven Deep Reinforcement Learning Framework

Here, we aim to improve the learning efficacy by proposing a new learning framework that exploits the efficiency of model-based optimization methods and the robustness of model-free ML approaches. The authors in [14] studied a similar concept of model-aided artificial intelligence in wireless systems. The model-based optimization is used to create a large set of *offline* training data for optimizing or refining the DNN models. This idea has been verified in IRS-assisted wireless systems to create training data sets for beamforming optimization, as in [7] and [9]. It works well, given either an accurate model or a tractable optimization solution. Different from [14], we used the DRL approach to build a robust outer-loop learning framework that is tolerable to uncertain information and system dynamics. We used the inner-loop optimization methods in the online learning phase to fast-track the control variables by solving approximate optimization problems efficiently. Comparing to [7, 9, 14], such an optimization-driven DRL framework can be applied to more complex wireless systems with both inaccurate models and intractable solutions. In the

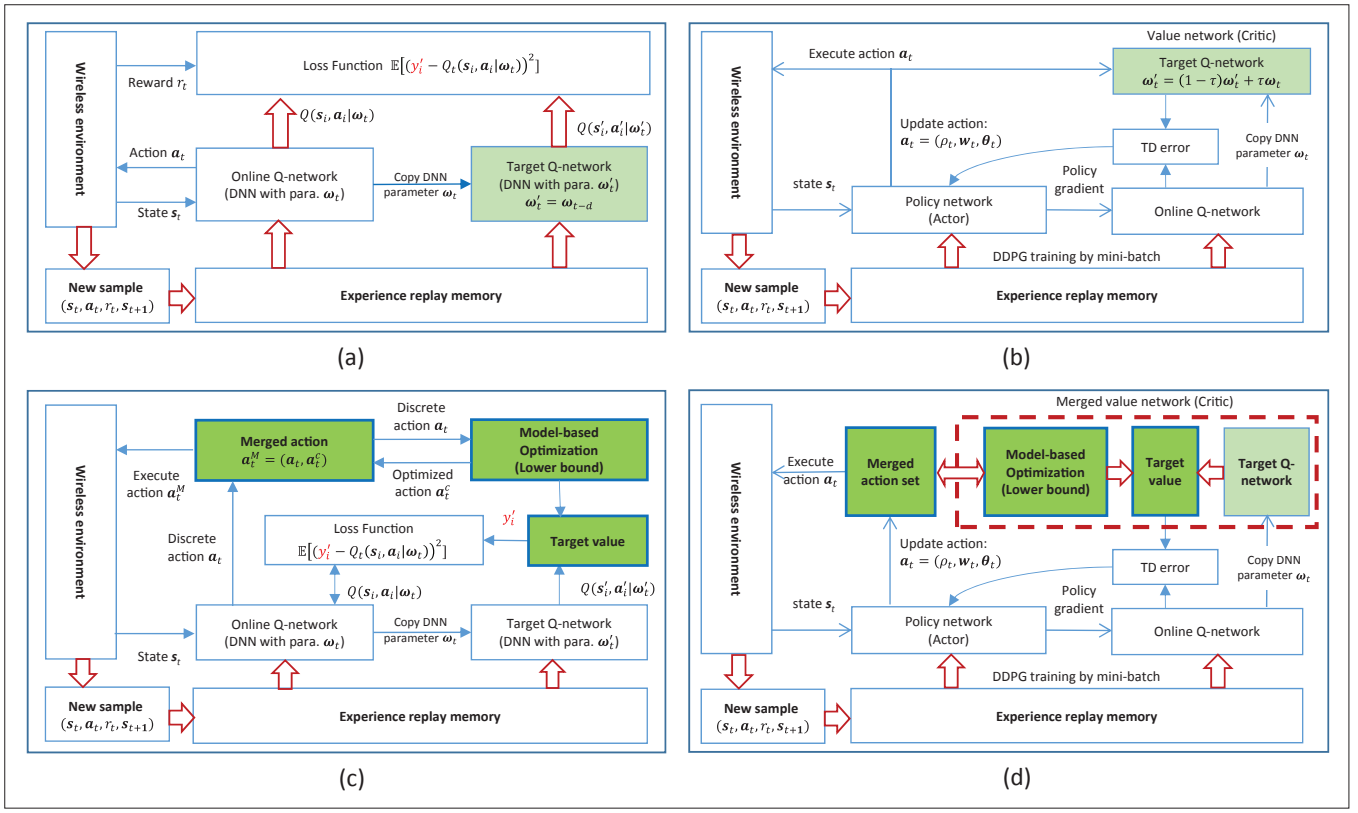

FIGURE 1. The comparison of model-free and optimization-driven DQN/DDPG algorithm: a) model-free DQN algorithm; b) model-free DDPG algorithm; c) optimization-driven DQN algorithm; d) optimization-driven DDPG algorithm.

sequel, we first analyzed the drawbacks of the conventional DRL approaches and then presented our optimization-driven DRL approach.

Conventional Model-Free DQN and DDPG Algorithms

DQN and DDPG are two conventional DRL algorithms that use DNNs to approximate the value and policy functions of the RL framework, respectively. To distinguish them from our algorithm design in this paper, we describe the conventional DQN and DDPG as "the model-free DRL algorithms" because they do not require exact system modeling and model-based optimization formulation. Generally, the model-free DQN algorithm is a natural extension of the traditional *Q*-learning method with finite action and state spaces. In wireless works, the model-free DQN can be used for optimizing the channel allocation, relay selection, user association, etc., which can be described by discrete control variables. It relies on the use of experience replay and target *Q*-network to stabilize the learning performance. The experience replay mechanism randomly selects a minibatch from a buffer of historical samples to train the DNN. As illustrated in Fig. 1a, the DNN training updates the DNN parameters of the online *Q*-network by minimizing the temporal difference error — that is, the mean squared difference between the online and the target *Q*-values. To stabilize the learning performance, DQN estimates the target *Q*-value by using a separate DNN, namely, the target *Q*-network, whose parameters are delayed copies of the online *Q*-network after a few decision epoches. The same idea also applies to the DDPG algorithm, which is an extension of the DQN algorithm to more complex control problems involving continuous variables (e.g., the RF transmit power and the IRS's phase shifts). As shown in Fig. 1b, the *Q*-value estimation in the DDPG algorithm is accompanied by a separate target Q-network, whose parameters are also evolving from the online *Q*-network.

Though the target *Q*-network in DQN or DDPG stabilizes the learning performance, the strong coupling between the online and target *Q*-networks may lead to a slow learning efficiency and reduced reward performance. First, both *Q*-networks can be randomly initialized and far from their optimum in the early stage of learning. This may mislead the learning process and require a large set of historical transition samples to ensure that the learning is correct. As such, the model-free DQN and DDPG practically require a long warm up period to train the online and target *Q*-networks. Second, it is problematic to configure the parameter copying from the online *Q*-network to the target *Q*-network. As shown in Fig. 1b, a small averaging parameter t in DDPG can stabilize but also slow down the learning process, while a large t implies strong correlation between the two Q-networks, resulting in performance fluctuations and even divergence.

Optimization-Driven DQN and DDPG Algorithms

To improve learning efficiency, we design the optimization-driven DRL framework that integrates the model-based optimization into the model-free DRL framework. We aim to stabilize and speed up the learning process by estimating the target *Q*-value in a better-informed and independent

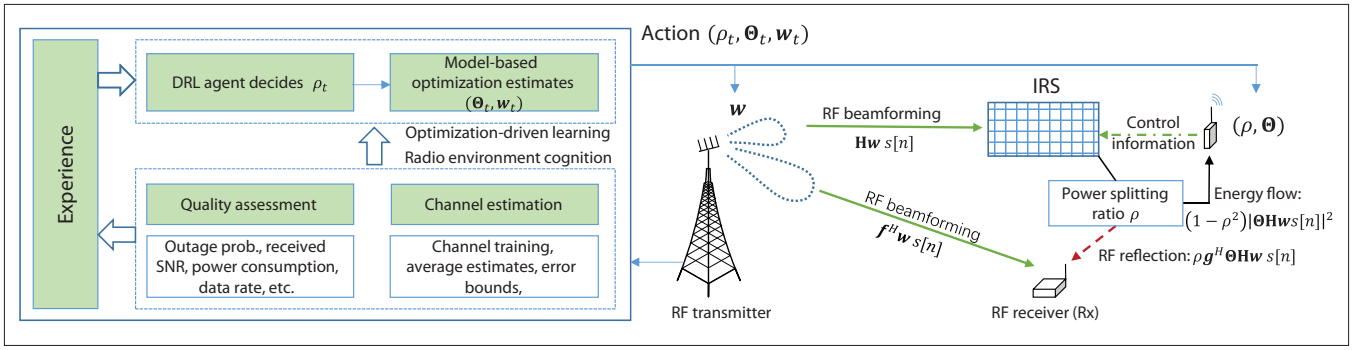

FIGURE 2. The application of optimization-driven DRL framework in IRS-assisted multi-input single output downlink system. Given the system information, the access point (AP) can search the optimal solution (ρ_t, θ_t, $\mathbf{w}_t$) at the beginning of each data transmission frame. The optimal solution is then executed by the AP and the IRS, respectively.

way. The motivation for the proposed framework stems from the following observations: The control variables of a complex problem are usually high dimensional and intractable by classical optimization methods. However, given a part of the control variables, the other part can easily be optimized in an approximate problem by exploiting the physical dependence among different control variables.

The main design principle is to split the control variables into two parts. One part can be obtained in the outer-loop DQN or DDPG with a reduced search space, while the other part can be optimized instantly and efficiently given the outer-loop control variables. Therefore, the splitting of control variables becomes a critical design problem. It determines the outer-loop learning performance as well as the computational complexity of the inner-loop optimization problem. One intuition for control variable splitting is to break the couplings among control variables and ensure that the inner-loop optimization is efficient and light-weight. To ensure robustness, the inner-loop optimization should be insensitive to the changes of system parameters (e.g., the channel state information) due to imprecise system modeling. In the sequel, we present two specific designs of the optimization-driven DRL.

Optimization-Driven DQN: For a mixed problem with both discrete and continuous decision variables, we can split the action vector into two parts — the discrete and continuous variables ($\mathbf{a}^t$, $\mathbf{a}_t^C$) as shown in Fig. 1c. The discrete actions can be updated by the model-free DQN. Given the discrete actions in the outer loop, we can resort to the optimization method to efficiently solve the continuous actions in a convex approximation of the original problem. The optimized actions can be combined with the discrete actions and then executed in the radio environment. One application of the optimization-driven DQN is the relay selection in an IRS-assisted wireless network. The outer-loop DQN can be used to pick the best relay node, while the inner-loop optimization determines the joint active and passive beamforming strategy given the relay selection strategy.

Optimization-Driven DDPG: For a high-dimensional continuous control problem, it is usually difficult to optimize all control variables in a single problem (e.g., due to non-convex problem structure). The model-free DQN algorithm also becomes inflexible as the continuous action and state spaces slow down the online learning performance. Similar to the optimization-driven DQN, we can follow the divide and conquer strategy to break the couplings among different control variables. In particular, we can learn a part of the control variables by the model-free DDPG algorithm, while solving the other part by optimization methods. As illustrated in Fig. 1d, when the outer-loop DDPG generates a part of the actions, an optimization module can be used to solve the other part of the action vector directly and provide a lower bound on the original problem. The splitting of the action vector should be properly designed so that the model-based optimization can be solved efficiently with a reduced computational complexity. For example, given the IRS's passive beamforming, the active beamforming can easily be optimized in an approximate convex problem. This implies that we can use outer-loop DDPG to search for the passive beamforming while optimizing the active beamforming in the inner loop.

From the above analysis, we highlight new features and novelties of the optimization-driven DRL framework:

- First, the optimization-driven DRL framework provides a general solution framework for complex problems that suffer from uncertain system dynamics, imprecise modeling, and high computational complexity. Compared to the model-free DRL methods, the optimization-driven DRL can reduce the search space and potentially improve the learning performance.
- Second, the control variable splitting offers a flexible trade-off between the efficiency of model-based optimization methods and the robustness of model-free DRL approaches. We can optimize the splitting of control variables according to resource constraints on computation capabilities, training overhead, and solution accuracy. In one extreme case, we can leave all control variables to the DRL approaches. This degenerates to the model-free DRL, which requires more training overhead. In the other case, all control variables can be solved by model-based optimization methods. However, this becomes inflexible and computational, requiring complex and dynamic systems.
- Third, we realize a novel *online integration* of model-based optimization and model-free DRL methods. Previous works typically use model-based optimization methods to simulate the data set for *offline* training [14],

whereas we integrate model-based optimization into the *online* decision-making cycles. Based on incomplete information, the inner-loop optimization provides a lower bound to the original problem, which can be used as a better-informed target Q-value for the outer-loop DNN training, especially in the early stage of learning. Besides, the optimization-driven target is independent of the online Q-network and can be more stable than the target Q-network. Such a decoupling between the online and target Q-networks can reduce the performance fluctuations and stabilize the learning faster.

To summarize, the optimization-driven DRL relies on the optimization module to guide its learning toward a better reward in a more stable and efficient way. Different from the supervised learning, the optimization-driven target is not given explicitly by the environment, but estimated by the DRL agent based on local observations in the online learning phase. Moreover, the optimization-driven DRL belongs to the DRL framework, which can be flexibly applied to different control problems with uncertain system dynamics, while the supervised learning is applicable to the cases when the testing data and the training data share similar features. That is, the features embedded in the training data can be extracted and applied to the testing data.

Case Study: Optimization-driven DDPG for Passive Beamforming

In this section, we examine the application of the optimization-driven DRL framework in an IRS-assisted multi-input single output downlink system. As illustrated in Fig. 2, the information transmissions from a multi-antenna access point (AP) to the receivers are assisted by the IRS with N reflecting elements. A few assumptions are listed as follows:

- The IRS sets a continuous phase shift and a flexible magnitude of reflection to reflect the incident RF signals. The extension to discrete phase shift is straightforward by limiting the feasible phase shift in a discrete set.
- The IRS is self-sustainable by harvesting RF energy. By controlling the magnitude of reflection, namely, the power-splitting (PS) ratio, a part of the incident signals is reflected to the receiver, while the other part can be absorbed by the IRS to sustain its operations.
- The channel information is uncertain due to estimation errors by the use of passive elements. The average channel estimates can be known by historic measurements, while the error estimates are randomly distributed within a convex and bounded set.

We aim to minimize the AP's transmit power by a joint beamforming optimization subject to the IRS's power budget and the receiver's SNR requirement similar to that in [15]. The control variables include the AP's active beamforming w, the IRS's PS ratio ρ, and the phase vector θ, as illustrated in Fig. 2. This problem is challenged by the non-convex coupling between the active and passive beamforming. The conventional alternative optimization method faces high computational complexity and becomes more difficult with uncertain channel information.

Different from the supervised learning, the optimization-driven target is not given explicitly by the environment, but estimated by the DRL agent based on local observations in the online learning phase.

Splitting Control Variables

We employ the optimization-driven DDPG to update the action $\mathbf{a}_t = (\rho_t, \mathbf{w}_t, \theta_t)$ in each decision epoch, which can be divided into two parts, the PS ratio ρ_t and two vectors $(\mathbf{w}_t, \theta_t)$. Given the PS ratio ρ_t in the outer-loop DDPG, we can easily determine a feasible phase vector θ_t and then solve the optimal active beamforming $\mathbf{w}_t$ efficiently in a convex optimization problem. As shown in Fig. 1d, the actor and critic networks of DDPG first generate the action and value estimates independently. Then, we fix the PS ratio ρ_t and feed it into the optimization module, which outputs the optimized solution $(\mathbf{w}_t, \theta_t)$ and also evaluates a lower bound on the target Q-value. If the optimization-driven target value is larger than the output of the target Q-network in the outer-loop DDPG algorithm, we can use it with a higher probability as the new target Q-value for DNN training; meanwhile, we update the action by the optimized solution. The splitting of control variables can also be performed in a different way. As the IRS's phase vector θ_t is difficult to optimize directly, we can search it by the outer-loop DDPG. Then, we can optimize the other variables $(\rho_t, \mathbf{w}_t)$ efficiently by a line search algorithm over ρ_t. Such a decomposition not only reduces the search space of the outer-loop DDPG, but also improves learning efficiency compared to the model-free DDPG.

Numerical Evaluation

The simulation follows the system model in Fig. 2. The AP-user distance in meters is $d_{\text{AP,User}} = 20$. The vertical distance from the IRS to the AP-user line segment is given by $d_{\text{IRS,User}} = 5$. Let $d_{\text{AP,IRS}}$ denote the horizontal distance from the AP to the IRS. The path loss at the unit distance is $L_0 = 30$ dB and the path-loss exponent equals 3.5, similar to [2]. The energy harvesting efficiency is $\eta = 0.5$. The noise power is −80 dBm. The reward of the outer-loop learning agent is defined as the ratio between the successfully transmitted data and the AP's total energy consumption. The size of IRS is $N = 20$ and the number of AP's antennas is $M = 2$. The DNN network structure and hyperparameters of the DDPG algorithms are listed in Table 2, where `fc(m, n)` denotes a fully connected neural network layer with the size of $m \times n$, `relu` and `sigmoid` denote different activation layers. The sizes of state and action spaces are specified by `N_STATES` and `N_ACTIONS`, respectively. In particular, for $M = 2$ and $N = 20$, the sizes of different control variables $(\theta_t, \mathbf{w}_t, \rho_t)$ can be easily determined as 40, 4 and 1, respectively. For fair comparison, we employ the same network structure for two DDPG algorithms. The difference between the optimization-driven and model-free DDPG algorithms lies in the size `N_ACTIONS` of action space. By the splitting of control variables, only a portion of the control variables are trained in the optimization-driven DDPG algorithm.

Better Reward Performance: Figure 3a demonstrates the dynamics of the AP's transmit power in the optimization-driven DDPG (denoted as the O-DDPG) compared with the conventional

Component	DNN structure	Hyperparameter	Value
Critic	fc(N_STATES, 64), fc(N_ACTIONS, 64)	batch size	32
	relu	learning rate (critic)	1e-4
	fc(64, 1)	learning rate (actor)	1e-3
Actor	fc(N_STATES, 64)	reward discount	0.5
	fc(64, N_ACTIONS)	memory capacity	1000
	sigmoid	replacement factor	0.01

TABLE 2. Network structure and hyperparameters in DDPG.

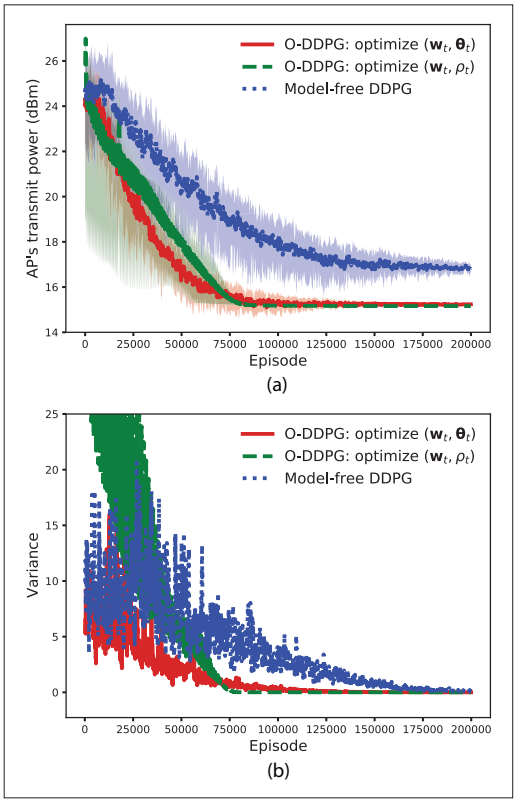

FIGURE 3. a) the AP's transmit power in different DDPG algorithms. The solid line denotes the median of 50 repetitions and the shaded regions in different colors cover 10th to 90th percentiles. b) the optimization-driven DDPG (O-DDPG) algorithms achieve more stable learning and faster convergence than the conventional model-free DDPG algorithm. The AP's and user's locations are set as $d_{\text{AP,User}}$ = 20, $d_{\text{AP,IRS}}$ = 10, $d_{\text{IRS,User}}$ = 5.

DDPG (denoted as the model-free DDPG). We consider different inner-loop optimization problems in the O-DDPG algorithms. In the first case, given the PS ratio rt learned by the outer-loop DDPG, we optimize ($\mathbf{w}_t$, θ_t) by a heuristic optimization method. In the second case, we search the passive beamforming θ_t in the outer-loop DDPG while optimizing (ρ_t, $\mathbf{w}_t$) by a bisection method. The inner-loop optimization solutions in both cases are used to guide the search for the optimal action (ρ_t, $\mathbf{w}_t$, θ_t). As the inner-loop optimizations generally provide better informed *Q*-value estimations, the O-DDPG algorithms can converge faster than the model-free DDPG and significantly decrease the AP's transmit power.

Faster and Stable Learning: In Fig. 3a, we also observe that the shaded areas of the O-DDPG algorithms are smaller than that of the model-free DDPG. This implies a more stable learning performance in the O-DDPG algorithms. We characterize the stability by the variance of reward over learning episodes. As shown in Fig. 3b, the rewards in two O-DDPG algorithms have smaller variances compared to that of the model-free DDPG. The convergence in the O-DDPG algorithms comes earlier, as the variance of reward gets closer to zero after a smaller number of learning episodes. It is clear that the O-DDPG algorithms stabilize after 10^6 episodes while the model-free DDPG requires 75 percent more episodes.

Optimal Deployment Location: Figure 4 shows the AP's transmit power when the IRS moves away from the AP to the receiver. The ideal situation is while the IRS's power demand is negligible, the IRS is always active and helpful to the AP's information transmission. Given a fixed SNR requirement at the receiver, the AP can reduce its transmit power significantly as the IRS moves closer to the receiver, as shown in Fig. 4a. This corroborates the previous observation in [2] that it is preferable to deploy the IRS closer to the receiver. However, in a more practical case with non-zero power demand at the IRS, the AP has to increase its transmit power to fulfill the IRS's power demand as it moves away from the AP, as shown in Fig. 4b. The reason is that a part of the AP's radio frequency (RF) power will be harvested by the IRS to sustain its operations. As the AP-IRS distance increases, the IRS will harvest a larger part of the AP's RF power by tuning the PS ratio, and thus contribute very little to the AP's information transmissions. This implies that the IRS should not be deployed far away from the RF transmitter, which is in contrast to the observation in [2] and reveals that the IRS's power demand becomes an important design aspect for the IRS's optimal deployment in wireless systems.

Challenges and Future Directions

Control Variable Splitting

The optimization-driven DRL divides the control variables into outer-loop learning and inner-loop optimization. This optimal division has to be carefully designed to achieve a balance between the computational complexity of optimization methods and the time efficiency of learning methods. Currently, there are no universal rules to guide the problem decomposition and control variable splitting. For the same control issue, we have different ways to split the control variables, which may lead to different learning performances. Hence, one of the future research directions is to study a systematic and analytical method to quantify the qualities of different splitting strategies.

Performance Trade-Off

The outer-loop learning has a reduced action space, which implies an increased learning speed with the potential cost of performance loss. The model-free DRL can flexibly explore the action space and efficiently exploit the best reward as learning continues. With a reduced action space, the optimization-driven DRL can achieve higher

convergence speed. However, the overall reward can also be affected, especially when a portion of the control variables are constrained by imprecise physical models and generally solved by the model-based optimization. As such, the learning efficiency and reward performance become two conflicting goals in the optimization-driven DRL. In our future work, the convergence performance can be further analyzed. Furthermore, it is worth exploring the trade-off between learning efficiency and reward performance, which will guide the optimal splitting of control variables.

Adaptive Integration

The optimization method improves the learning efficiency significantly in the initial stage while contributing little as the reward increases. In the initial stage, the learning process intends to explore random actions without sufficient training samples from the past experiences. As such, the optimized action will be more often selected by the DRL agent as it provides a relatively better reward. This implies an adaptive integration of the optimization and learning methods during the system's evolution. We expect that it is not necessary to perform frequent and computation-intensive optimization in each learning episode. Hence, the adaptive integration allows the DRL agent to execute the optimization module on demand, such as only when the outer-loop variables or the reward performance have been changed significantly.

Optimization-Driven Multi-Agent Learning

This work shows the feasibility and benefits of an integration between model-based optimization and the model-free DRL. The design of other optimization-driven ML approaches, or a combination of them, is worth further investigation. For example, we can extend the optimization-driven DRL concept to multi-agent systems, in which an individual agent makes its own decision based on the observations and actions of the other agents. This implies a large set of training data and slow convergence if we employ the conventional multi-agent DRL methods. Instead, by solving model-based optimization locally, each agent can estimate the optimal actions of the other agents, which may help guide its learning toward a better reward with enhanced learning efficiency.

Application to Nonlinear Backscatter Channel Estimation

The Internet-of-Things (IoT) requires ultra-low power communication, such as backscatter communication, as IoT devices should be operated with limited power and complexity for self-sustainability. The recent effort has been directed to deploying wireless-powered backscatter communication (WPBC) networks. In this context, the energy beamforming for wireless energy transfer to IoT devices and the receiver beamforming for wireless information transmission are the prerequisites for effective deployment of WPBC, for which the channel estimation is challenging because of the nonlinear backscatter channels. The existing LS/MMSE methods for backscatter channel estimation are not optimal, so that we may apply the optimization-driven ML method for such nonlinear backscatter channel estimation, in conjunction with the LS/MMSE methods for enhanced learning efficiency.

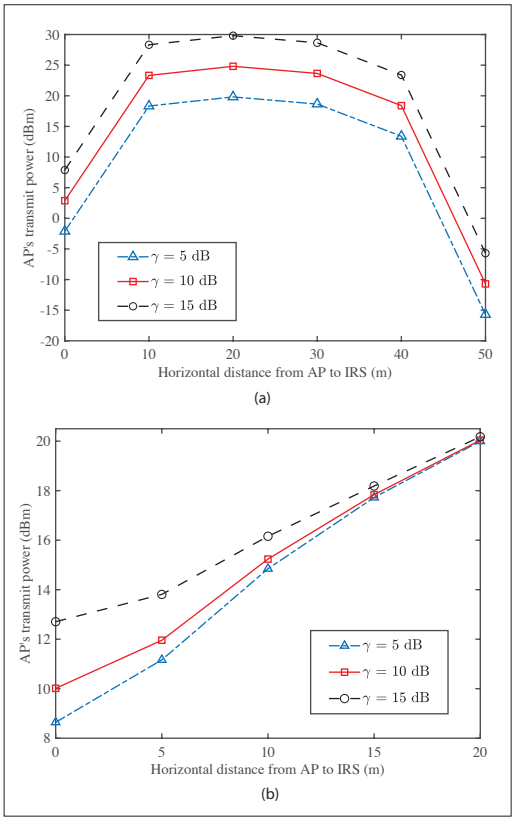

FIGURE 4. The AP's transmit power varies as the IRS moves from the AP to the receiver: a) the IRS's power demand is neglectable in the ideal case. The AP-User distance is $d_{AP,User}$ = 50 meters; b) the IRS has the constant power demand of 20 μW, which has to be fulfilled by RF energy harvesting. The AP-User distance $d_{AP,User}$ = 20 meters.

Conclusion

In this article, we have reviewed the applications of ML approaches in IRS-assisted systems. An inspection of the common limitations of existing ML approaches has motivated us to design a novel optimization-driven DRL framework for the joint beamforming optimization problem. Numerical results have demonstrated that the proposed approach improves the learning efficiency and reward performance significantly compared to the conventional model-free DRL methods.

Acknowledgments

The work of Shimin Gong was supported in part by the Shenzhen Fundamental Research Program under Grant JCYJ20190807154009444 and the National Natural Science Foundation of China (NSFC) under grant number 61972434. The work of Dusit Niyato was supported in part by the National Research Foundation, Singapore, under its Campus for Research Excellence and Technological Enterprise (CREATE) program; Alibaba Group through Alibaba Innovative Research (AIR) program and Alibaba-NTU Singapore Joint Research Institute, and the National Research Foundation, Singapore under the AI Singapore program (AISG) (AISG2-RP-2020-019), WASP/NTU grant number M4082187 (4080), Singapore Ministry of Education (MOE) Tier 1 (RG16/20).

The work of Dong In Kim was supported in part by the National Research Foundation of Korea grant funded by the Korean government (MSIT), grant number 2021R1A2C2007638.

References

[1] S. Gong *et al.*, "Toward Smart Wireless Communications via Intelligent Reflecting Surfaces: A Contemporary Survey," *IEEE Commun. Surv. Tut.*, June 2020, pp. 1–33.
[2] Q. Wu and R. Zhang, "Intelligent Reflecting Surface Enhanced Wireless Network via Joint Active and Passive Beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, Nov. 2019, pp. 5394–5409.
[3] X. Yu, D. Xu, and R. Schober, "Enabling Secure Wireless Communications via Intelligent Reflecting Surfaces," *Proc. IEEE GLOBECOM*, Dec. 2019, pp. 1–6.
[4] S. Y. Park and D. I. Kim, "Intelligent Reflecting Surface-Aided Phase-Shift Backscatter Communication," *Proc. Int. Conf. Ubiquit. Inf. Manag. Commun. (IMCOM)*, Jan. 2020, pp. 1–5.
[5] A. M. Elbir *et al.*, "Deep Channel Learning for Large Intelligent Surfaces Aided Mm-Wave Massive MIMO Systems," *IEEE Wireless Commun. Lett.*, May 2020, pp. 1–5.
[6] S. Liu *et al.*, "Deep Denoising Neural Network Assisted Compressive Channel Estimation for mmWave Intelligent Reflecting Surfaces," *IEEE Trans. Veh. Techn.*, vol. 69, no. 8, Aug. 2020, pp. 9223–28.
[7] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Deep Learning for Large Intelligent Surfaces in Millimeter Wave and Massive MIMO Systems," *Proc. IEEE GLOBECOM*, Dec. 2019, pp. 1–6.
[8] T. Jiang, H. V. Cheng, and W. Yu, "Learning to Reflect and to Beamform for Intelligent Reflecting Surface With Implicit Channel Estimation," *IEEE JSAC*, vol. 39, no. 7, July 2021, pp. 1931–45.
[9] J. Gao *et al.*, "Unsupervised Learning for Passive Beamforming," *IEEE Commun. Lett.*, vol. 24, no. 5, May 2020, pp. 1052–56.
[10] G. Lee *et al.*, "Deep Reinforcement Learning for Energy-Efficient Networking With Reconfigurable Intelligent Surfaces," *Proc. IEEE ICC*, Jul. 2020, pp. 1–6.
[11] K. Feng *et al.*, "Deep Reinforcement Learning Based Intelligent Reflecting Surface Optimization for MISO Communication Systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, Jan. 2020, pp. 745–49.
[12] C. Huang, R. Mo, and C. Yuen, "Reconfigurable Intelligent Surface Assisted Multiuser MISO Systems Exploiting Deep Reinforcement Learning," *IEEE JSAC*, vol. 38, no. 8, Aug. 2020, pp. 1839–50.
[13] H. Yang *et al.*, "Deep Reinforcement Learning Based Intelligent Reflecting Surface for Secure Wireless Communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, Jan. 2021, pp. 1963–74.
[14] A. Zappone *et al.*, "Model-Aided Wireless Artificial Intelligence: Embedding Expert Knowledge in Deep Neural Networks for Wireless System Optimization," *IEEE Veh. Techn. Mag.*, vol. 14, no. 3, Sept. 2019, pp. 60–69.
[15] J. Lin *et al.*, "Deep Reinforcement Learning for Robust Beamforming in IRS-Assisted Wireless Communications," *Proc. IEEE GLOBECOM*, Dec. 2020, pp. 1–6.

Biographies

Shimin Gong [M'15] is an associate professor at the School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou, China. His research interests include wireless powered communications and machine learning in wireless communications. He was a recipient of the Best Paper Award on MAC and Cross-Layer Design in IEEE WCNC 2019.

Jiaye Lin is pursuing his bachelors degree at the School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou, China. His research interests include The Internet of Things, wireless power transfer, and backscatter communications.

Beichen Ding received a Ph.D in mechanical engineering from the University of Bath, UK. He is an assistant professor at the School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou, China. His current research interests include The Industrial Internet of Things, sensors and motion control of smart machine systems, and adaptive control for nonlinear systems.

Dusit Niyato [M'09, F'17] is a professor at the School of Computer Science and Engineering at Nanyang Technological University, Singapore. He received his Ph.D in electrical and computer engineering from the University of Manitoba, Canada. His research interests include energy harvesting for wireless communication, The Internet of Things, and sensor networks.

Dong In Kim [F'19] received his Ph.D in electrical engineering from the University of Southern California, Los Angeles. He is a SKKU Fellowship professor with the College of Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea. He is a Fellow of the Korean Academy of Science and Technology, and a member of the National Academy of Engineering of Korea. He was selected the 2019 recipient of the IEEE Communications Society Joseph LoCicero Award for Exemplary Service to Publications. He is the general chair for IEEE ICC 2022 in Seoul, South Korea.

Mohsen Guizani [M'89, CSM'99, CF'09] is a professor at Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE. Previously, he worked in different institutions in the USA. His research interests include applied machine learning and artificial intelligence, Internet of Things (IoT), intelligent systems, smart city, and cybersecurity. He was elevated to IEEE Fellow in 2009 and was listed as a Clarivate Analytics Highly Cited Researcher in Computer Science in 2019, 2020 and 2021. He has won several research awards including the "2015 IEEE Communications Society Best Survey Paper Award," the Best ComSoc Journal Paper Award in 2021 as well five Best Paper Awards from ICC and Globecom Conferences.