Deep Learning based Wireless Resource Allocation with Application to Vehicular Networks

Le Liang, Member, IEEE, Hao Ye, Student Member, IEEE, Guanding Yu, Member, IEEE, and Geoffrey Ye Li Fellow, IEEE

Abstract—It has been a long-held belief that judicious resource allocation is critical to mitigating interference, improving network efficiency, and ultimately optimizing wireless communication performance. The traditional wisdom is to explicitly formulate resource allocation as an optimization problem and then exploit mathematical programming to solve the problem to a certain level of optimality. Nonetheless, as wireless networks become increasingly diverse and complex, e.g., the high-mobility vehicular networks, the current design methodologies face significant challenges and thus call for rethinking of the traditional design philosophy. Meanwhile, deep learning, with many success stories in various disciplines, represents a promising alternative due to its remarkable power to leverage data for problem solving. In this paper, we discuss the key motivations and roadblocks of using deep learning for wireless resource allocation with application to vehicular networks. We review major recent studies that mobilize the deep learning philosophy in wireless resource allocation and achieve impressive results. We first discuss deep learning assisted optimization for resource allocation. We then highlight the deep reinforcement learning approach to address resource allocation problems that are difficult to handle in the traditional optimization framework. We also identify some research directions that deserve further investigation.

Index Terms—Deep learning, reinforcement learning, resource allocation, wireless communications, vehicular networks

I. INTRODUCTION

OVER the past few decades, wireless communications have been relentlessly pursuing higher throughput, lower latency, higher reliability, and better coverage. In addition to designing more efficient coding, modulation, channel estimation, equalization, and detection/decoding schemes, optimizing the allocation of limited communication resources is another effective approach [1].

From Shannon’s information capacity theorem [2], power and bandwidth are the two primitive resources in a communication system. They determine the capacity of the wireless channel, up to which the information can be transmitted with an arbitrarily small error rate. For modern wireless communication systems, the definition of communication resources has been substantially enriched. Beams in a multiple-input multiple-output (MIMO) system, time slots in a time-division multiple access system (TDMA), frequency sub-bands in a frequency-division multiple access system, spreading codes in a code-division multiple access system, and even base stations or backhaul links in virtualized wireless networks all count. Judicious allocation of these communication resources in response to channel conditions and user’s quality-of-service requirements is critical in wireless system optimization. For example, water-filling power allocation needs to be performed over different subcarriers in an orthogonal frequency-division multiplexing (OFDM) system or different channel eigen directions in a MIMO system for capacity maximization. Power and spectrum allocation is important in a cellular network with device-to-device (D2D) underlay to manage interference and optimize network throughput. In addition to optimizing for traditional physical layer communication metrics, such as capacity maximization [3] and power minimization [4], cross-layer resource allocation takes account of the requirements of upper layers, e.g., delay and fairness, through optimizing properly defined utility functions [5]–[7].

Historically, the dominant approach to resource allocation is through mathematical programming, where we optimize one of the design criteria of interest, e.g., sum rate maximization or interference minimization, while imposing bounds on the remaining. Despite the remarkable success in this domain, it turns out many of the formulated optimization problems are difficult to solve [8]. Moreover, with a myriad of new applications to support, conventional methods find it increasingly difficult to balance and model the diverse service requirements in a mathematically exact way. As an example, for ultra-reliable and low-latency communications (URLLC), one of the three major 5G usage scenarios, the definitions of latency and reliability are still subject to debate [9], not to mention a principled approach to provide performance guarantee. A more flexible framework for wireless resource allocation is thus needed, motivating a departure from the traditional wireless design philosophy.

Recently, adaptation of machine learning tools to address difficult problems in wireless communications and networking has gained momentum, ranging from physical layer design [10]–[15], resource allocation [16], [17], networking [18], caching [19], to edge computing [20]. In fact, the latest cycle of enthusiasm of machine learning is largely triggered by the exceptional performance of deep learning in a broad array of application scenarios. Deep learning enables a powerful data-driven approach to many problems that were deemed hard due to, e.g., lack of accurate models or prohibitive computational complexity. In the wireless resource allocation context, deep learning has been shown to achieve significant performance improvement over conventional methods in several recent works [21]–[24]. In particular, it has been demonstrated that
deep reinforcement learning (RL) is capable of providing a nice treatment of service requirements that are hard to model exactly and thus also not subject to any effective optimization approaches. The goal of this paper is to review some of the most promising results and discuss the principles, benefits, and potential challenges of leveraging deep learning to address wireless resource allocation problems in general with application to vehicular networks as a special example. Since this is a fast evolving field, we do not attempt to exhaustively cover all research papers in this area but only highlight those that closely align with our theme. We refer interested readers to other excellent survey and tutorial papers on various aspects of leveraging learning concepts in the wireless context [18], [19], [25]–[30] to get a complete picture. Compared to them, this paper differentiates itself in that we exclusively focus on addressing wireless resource allocation using deep learning. We emphasize the fundamental properties of this category of problems that make the data-driven deep learning approach appealing and demonstrate through extensive example studies how to unleash the power of this promising method to its fullest extent.

The paper is organized as follows. In Section II, we discuss the limitations of traditional optimization methods for wireless resource allocation and motivate deep learning in addressing the issue. In Section III, we present examples on how to leverage deep learning to solve resource optimization problems in a more efficient way. In Section IV, deep RL based methods are discussed in detail that warrant a fundamental shift in treating resource allocation problems in a more flexible and effective framework. In Section V, we recognize and highlight several open issues that are worth further research. Concluding remarks are finally made in Section VI.

II. MACHINE LEARNING FOR RESOURCE ALLOCATION

![Fig. 1. Classification of approaches to wireless resource allocation.](image)

As in Fig. 1, the mainstream approach to wireless resource allocation has long been to formulate the design objective and constraints as an optimization problem. It is then solved optimally or sub-optimally, depending on problem complexity and allowable computation time, by leveraging tools from various disciplines, including mathematical programming, graph theory, game theory, etc. In this section, we discuss the limitations of these optimization approaches and highlight the potentials of uprising data-driven approaches enabled by machine learning, in particular deep learning.

A. Limitation of Traditional Optimization Approach

Except in a few simple cases, where we are fortunate enough to end up with convex optimization that admits a systematic procedure to find the global optimum, most optimization problems formulated for wireless resource allocation are strongly non-convex (continuous power control), combinatorial (discrete channel assignment), or mixed integer nonlinear programming (combined power control and spectrum assignment). For instance, it has been shown in [8] that the spectrum management problem in a frequency selective channel, where multiple users share the same spectrum, is non-convex and NP-hard. There are no known algorithms that can solve the problem to optimality with polynomial time complexity. To deal with problems of this kind, we are often satisfied with a locally optimal solution or some good heuristics without performance guarantee. More often than not, even these “sub-optimal” methods are computationally complex and hard to be executed in real time. Another limitation of existing optimization approaches points to the requirement of exact solutions, which tends to abstract away many imperfections in reality for mathematical tractability, and the solution is highly dependent on the accuracy of models. However, the wireless communication environment is constantly changing by nature and the resultant uncertainty in model parameters, e.g., channel information accuracy, undermines the performance of the optimized solution.

Finally, as wireless networks grow more complex and versatile, a lot of the new service requirements do not directly translate to the performance metrics, such as sum rate or proportional fairness, that the communication community is used to. For example, in high-mobility vehicular networks, the simultaneous requirements of capacity maximization for vehicle-to-infrastructure (V2I) links and reliability enhancement for vehicle-to-vehicle (V2V) links [27], [28] do not admit an obvious formulation. In particular, if we define the reliability of V2V links as the successful delivery of packets of size $B$ with the time constraint $T_0$ [31], [32], the problem becomes a sequential decision problem spanning the whole $T$ time steps and is difficult to solve in a mathematically exact manner. To avoid such difficulties, traditional optimization based methods break down the problem into isolated resource allocation decisions at each time step without considering the long-term effect. For instance, methods in [35]–[40] reformulate the requirement as a signal-to-interference-plus-noise ratio (SINR) constraint at each decision step and then use various optimization techniques to solve for the resource allocation solution. Such a practice loses the flexibility to balance V2I and V2V performance across the whole time $T$ and leads to inevitable performance loss.

B. Deep Learning Assisted Optimization

Deep learning allows multi-layer computation models that learn data representation with multiple levels of abstraction
It has seen a recent surge in a wide variety of research areas due to its exceptional performance in many tasks, such as speech recognition and object detection. Coupled with availability of more computing power and advanced training techniques, deep learning enables a powerful data-driven approach to many problems that are deemed difficult traditionally. In the context of wireless resource allocation, this sheds light on (partially) solving hard optimization problems.

In the simplistic form, deep learning can be leveraged to learn the correspondence of the parameters and solutions of an optimization problem. The computationally complex procedure to find optimal or sub-optimal solutions can be taken offline. With the universal approximation capability of deep neural networks (DNNs), the relation between the parameter input and the optimization solution obtained from some existing algorithms can be approximated. For implementation in real time, the new parameter is input into the trained DNN and a good solution can be given almost instantly, thus improving its potential for adoption in practice.

In learning tasks, the DNN is usually trained to minimize the discrepancy between the output and the ground truth given an input. With this idea in mind, we can leverage deep learning to directly minimize or maximize the optimization objective, i.e., treating the objective as the loss function in supervised learning. Then various training algorithms, such as stochastic gradient descent, can be employed to find the optimization solution. Compared with the direct input-output relation learning, this approach goes beyond the performance limit imposed by the traditional optimization algorithms that are used to generate the training data. Alternatively, deep learning can be embedded as a component to accelerate some steps of a well-behaved optimization algorithm, such as the pruning stage of the branch-and-bound in [39], [40]. This method approaches the existing algorithm that is known to achieve near-optimal performance but significantly reduces its execution time.

C. Deep Reinforcement Learning based Resource Allocation

RL concerns sequential decision making so as to maximize a numeric reward signal through interacting with the unknown environment, as illustrated in Fig. 2 Mathematically, the RL problem can be modeled as a Markov decision process (MDP). At each discrete time step \( t \), the agent observes some representation of environment state \( S_t \) from state space \( \mathcal{S} \), and then selects an action \( A_t \) from action set \( \mathcal{A} \). Following the action, the agent receives a numerical reward \( R_{t+1} \) and the environment transitions to a new state \( S_{t+1} \), with transition probability \( p(s', r|s, a) \). In RL, decision making manifests itself in a policy \( \pi(a|s) \), which is a mapping from states in \( \mathcal{S} \) to probabilities of selecting each action in \( \mathcal{A} \). The goal of learning is to find an optimal policy \( \pi^* \), that maximizes the expected accumulative rewards from any initial state \( s \).

The RL framework provides native support for addressing sequential decision making under uncertainty that we encounter in, e.g., the resource allocation problem in vehicular networks. The problem can be tackled by designing reward signal that correlates with the ultimate objective and the learning algorithm can figure out a decent solution to the problem automatically. Indeed, the flexibility of reward design in RL provides a solution to many problems that are hard to model in a mathematically exact manner. In addition, distributed algorithms are made possible by RL approaches, where each wireless node can be treated as an agent and learn to allocate resources efficiently. It eliminates the need to collect global channel state information (CSI) that incurs heavy signaling overhead and inevitable delay-caused uncertainty.

A rich set of algorithms, including SARSA, Q-learning, policy gradient, Actor-Critic, the methods integrating planning and learning, such as Dyna, etc. [41], have been developed to provide efficient sampling based methods to solve RL problems. They can also be combined with the latest advances in deep learning to bring the performance to a new level, such as deep Q-learning that has achieved remarkable performance in human-level video game play [42].

1) Q-Learning: Q-Learning [43] is a popular model-free method to solve RL problems. It is based on the concept of action-value function, \( q_\pi(s, a) \), for policy \( \pi \), which is defined as the expected return starting from state \( s \) taking action \( a \), and then following policy \( \pi \). The action-value function of the optimal policy, \( q_\star(s, a) \), satisfies a recursive relation, known as the Bellman optimality equation. In principle, one can solve this system of nonlinear equations for \( q_\star(s, a) \) if the dynamics \( p(s', r|s, a) \) are known. Once \( q_\star \) is obtained, it is easy to determine the optimal policy. Q-learning avoids the difficulties of acquiring the dynamics by taking an iterative update approach, given by

\[
Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right],
\]

where \( \alpha \) is the step-size parameter, \( \gamma \in (0, 1] \) is the MDP discounter factor, and the choice of \( A_t \) in state \( S_t \) follows some soft policies, e.g., the \( \epsilon \)-greedy, meaning that the action

![Fig. 2. The agent-environment interaction in reinforcement learning.](image-url)

\[
Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right],
\]

where \( \alpha \) is the step-size parameter, \( \gamma \in (0, 1] \) is the MDP discounter factor, and the choice of \( A_t \) in state \( S_t \) follows some soft policies, e.g., the \( \epsilon \)-greedy, meaning that the action
with maximal estimated value is chosen with probability $1 - \epsilon$ while a random action is selected with probability $\epsilon$.

2) **Deep Q-Network with Experience Replay:** In many problems of practical interest, the state and action space can be too large to store all action-value functions in a tabular form. As a result, it is common to use function approximation to estimate these value functions. In deep Q-learning [42], a DNN parameterized by $\theta$, called deep Q-network (DQN), is used to represent the action-value function. The state-action space is explored with some soft policies, e.g., $\epsilon$-greedy, and the transition tuple $(S_t, A_t, R_{t+1}, S_{t+1})$ is stored in a replay memory at each time step. The replay memory accumulates experiences over many episodes of the MDP. At each step, a mini-batch of experiences $D$ are uniformly sampled from the memory for updating $\theta$ with variants of stochastic gradient-descent methods, hence the name experience replay, to minimize the sum-squared error:

$$\sum_{D} \left[ R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a'; \theta^{-}) - Q(S_t, A_t; \theta) \right]^2,$$

(2)

where $\theta^{-}$ are the parameters of a target Q-network, which are duplicated from the training Q-network parameters $\theta$ periodically and fixed for a couple of updates. Experience replay improves sample efficiency through repeatedly sampling stored experiences and breaks correlation in successive updates, thus also stabilizing learning.

In fact, there are a good number of existing studies leveraging the concept of MDP for resource allocation, e.g., the delay-optimal OFDMA power control and subband allocation in [44], [45] and virtualized radio resource scheduling for software-defined vehicular networks in [46]. However, we do not treat them as (deep) RL approaches as they assume the availability of MDP transition dynamics more or less and are not learning from interactions with unknown environment. Such trial-and-error learning behavior is a key ingredient in making RL as successful as today, from our perspective.

### III. Deep Learning Assisted Optimization for Resource Allocation

This section deals with the employment of deep learning to find (near-) optimal solutions of the optimization problems for wireless resource allocation in an efficient manner. There are three ways to incorporate deep learning in solving the optimization problems:

- The supervised learning paradigm: the DNNs are applied to learn the input-output relation of the solution of a given algorithm.
- The objective-oriented unsupervised learning paradigm: the optimization objective is employed as the loss function, which is optimized directly during training.
- The optimization guided paradigm: the structure of the optimization problem is used to improve sample efficiency or robustness of the learning algorithms.

#### A. Supervised Approaches for Optimization

The most straightforward way to leverage deep learning for resource allocation problems is treating a given optimization problem as a black box and using various deep learning techniques to learn its input/output relation. In this case, a traditional optimization method will act as a supervisor, whose output will serve as the ground truth for training the DNNs. With the universal approximation ability of DNNs, the mapping between the input and solution of the traditional optimization method can be approximated.

The training and testing stages are conveniently shown in Fig. 3. During the training stage in Fig. 3(a), labeled samples are obtained by running the mathematical algorithm of the involved optimization problem using simulated data. All generated samples can be split randomly into training and validation sets. We use the training set to minimize the errors by updating the weights of DNNs. In the testing stage in Fig. 3(b), we also generate the labeled testing samples using the mathematical algorithm. We then pass the input to the trained network and collect the inferred solution before comparing it with its corresponding true label.

![Fig. 3. The supervised paradigm of using deep learning for optimization.](image)

In prior works, DNNs have shown their power in mimicking the solutions of state-of-the-art heuristic algorithms while reducing computational complexity. In [47], the non-convex power allocation problem is NP-hard and DNNs are used to approximate the state-of-art heuristic algorithm, i.e., the weighted minimum mean square error (WMMSE) algorithm. The simulation results suggest that the performance of DNN is very close to that of the WMMSE algorithm while significantly reducing computational complexity.

In [48], a similar paradigm is used to deal with a classical combinatorial optimization problem, i.e., the linear sum assignment programming (LSAP) problem, which is widely used in wireless resource allocation. LSAP problem is about how to assign $n$ jobs to $n$ people in a best way so that some utility (cost) function can be maximized (minimized). The traditional method for the LSAP problem is the Hungarian algorithm [49] with the computational complexity of $O(n^3)$, which is considered to be impractical for many applications. To reduce complexity, the LSAP problem has been first decomposed...
into several sub-assignment problems, which are classification problems. Then, DNNs have been utilized to solve each sub-assignment problem. Finally, the low-complexity greedy collision-avoidance rule has been used to get the inferred output for the LSAP problem. As shown in Table I, DNNs, including the feed-forward network (FNN) and convolutional neural network (CNN), can be used to obtain a real-time solution to the LSAP problem with slight loss of accuracy.

In addition, the features of the communication network could be exploited in order to reduce the number of required training samples. For instance, a link scheduling method without CSI has been proposed in [50] and [51] for D2D networks with the help of feature embedded paradigm. The link scheduling problem focuses on a densely deployed D2D network with a large number of mutually interfering links. The goal of link scheduling problem is to maximize the network utility by activating a subset of links at any given time. It can be formulated as a non-convex combinatorial optimization problem. Traditional methods are based on different mathematical optimization techniques with the help of accurate CSI. For a network with \( N \) D2D links, \( N^2 \) channel coefficients need to be estimated. Therefore, channel estimation stage is time- and resource-consuming. In order to bypass the channel estimation stage, the feature embedded paradigm is used for link scheduling. In [50], transmitter and receiver density grids are first constructed for a given D2D network, and then two designed convolutional filters are used to learn the interference pattern among different D2D links. The convolutional stage corresponds to the feature extraction process in Fig. 4. The results of the convolutional stage with other features are input into a DNN to learn the scheduling result in an unsupervised manner. The simulation results suggest that DNNs can effectively learn the network interference topology and perform scheduling to near optimum without the help of accurate CSI. The proposed method needs 800,000 training samples and has good scalability and generalizability to different topologies.

To further reduce the number of required training samples, a graph embedding method has been used in [51] to achieve this goal. For a given D2D network, a graphical model is constructed, where each node corresponds to a D2D link. Then graph embedding is used to learn the vector representation of each node based on the topology information. The output feature vector has then been input to a DNN for link scheduling. Compared with the kernel based feature extraction process in [50], the graph embedding process only involves the nonlinear function mapping and can be learned with fewer training samples. The simulation results are summarized in Table II.

The supervised paradigm suffers from several disadvantages. First, since the ground truth must be provided by conventional optimization approaches, performance of the deep learning system will be bounded by that of the used conventional approaches. Second, a large amount of training samples are required to get a good model. For example, 1,000,000 and 50,000 training samples are used in [47] and [48], respectively. However, labeled training samples are always difficult to obtain, especially in wireless communications. Third, the scalability of the learned model using black box paradigm is limited, which means that the output performance deteriorates sharply with the complexity of the problems.

**B. Unsupervised Approaches for Optimization**

In order to improve performance of the deep learning enabled optimization, unsupervised approaches have been proposed to train the DNN according to the optimization objective directly, instead of learning from the conventional optimization approach. In general, deep learning is trained to optimize a loss function via stochastic gradient descent. The loss functions are designed from case to case. For instance, in classification problems, the cross-entropy loss is often used while in regression problems, \( l_1 \) and \( l_2 \) losses are preferred. Therefore, it is natural to use the objective function from the optimization problem as the loss function so that the objective function can be optimized based on the stochastic gradient descent during training.

Various optimization loss functions have been utilized to train the DNNs for wireless resource management and perform better than the state-of-the-art heuristics. In [53], the CNN is used for a power control strategy in wireless communications.

![Figure 4](image.png)

**Table I**

| Problem | Mathematical Algorithm | Optimal Solution |
|---------|------------------------|------------------|
| Instance | | |

**Table II**

| Number of Training Samples | 200 | 1000 | 1500 |
|----------------------------|-----|------|------|
| **Classifier Accuracy**     | 0.8120 | 0.8192 | 0.8388 |
| **Average Sum Rate**       | 0.9362 | 0.9406 | 0.9447 |

**Note:**

| Parameter | Value |
|-----------|-------|
| Time(s)   | 0.5916 |
| Accuracy  | 100%   |
| Time(s)   | 0.0120 |
| Accuracy  | 92.76% |
| Time(s)   | 0.0040 |
| Accuracy  | 90.80% |

**Fig. 4**

Training stage for the feature embedded paradigm.

| Number of Training Samples | 200 | 500 | 1000 | 1500 |
|----------------------------|-----|-----|------|------|
| **Classifier Accuracy**    | 0.8120 | 0.8208 | 0.8392 | 0.8388 |
| **Average Sum Rate**       | 0.9362 | 0.9395 | 0.9406 | 0.9447 |
The system can be trained to optimize the spectral efficiency or the energy efficiency, which can be expressed as

$$\eta_{SE} = \sum_{i \in I} \log_2 \left( 1 + \frac{h_{i,i}P_i}{N_0W + \sum_{k \in I \setminus \{i\}} h_{k,i}P_k} \right),$$

and

$$\eta_{EE} = \sum_{i \in I} \frac{\eta^{(i)}_{SE}}{P_i + P_e},$$

respectively, where $P_i$ and $P_e$ denote the transmit power of transmitter $i \in I$ and the circuit power, respectively. $N_0$ is noise spectral density, $W$ is the bandwidth, $h_{i,j}$ denotes the channel from transmitter $i \in I$ to receiver $j \in J$. The performance is shown to be better than the conventional WMMSE approach.

Although training DNNs is a non-convex optimization problem, the loss functions used are often convex. The convex loss functions bring benefits to the DNN training procedure [54]. In the resource allocation problems, however, the loss functions are often non-convex. Therefore, issues arise from the theoretical and practical aspects in obtaining good performance. In previous works, additional techniques have been tried to improve the performance of training with the optimization objective in the wireless resource allocation framework. In [53], the network is pre-trained with the WMMSE solution as the ground truth. In [55], multiple deep networks are ensembled together for better performance. In [50], special structures of the networks are designed, including convolutional filters and feedback links.

C. Optimization Guided Deep Learning

In the two aforementioned learning paradigms, the optimization procedure is viewed as a black box and completely replaced by a deep learning model. These approaches do not require any prior information about the optimization problems, but need a large amount of training data in order to get good performance. In addition, they are suitable for optimization problems with only one kind of output variables. For example, the power allocation problem in [47] only includes continuous output variables, whereas combinatorial ones are involved in [48], [50], [51]. They find difficulties in dealing with optimization problems with different types of output [39] or with stochastic constraints [56]. In order to overcome these shortcomings, structures of the optimization problem can be employed to assist deep learning.

The mixed integer nonlinear programming (MINLP) problems are frequently encountered in wireless resource allocation, including user association, subcarrier allocation, and computation offloading. MINLP problems are generally NP-hard and difficult to obtain the optimal solutions. The general formulation is given by

$$\begin{align*}
\text{maximize} & \quad f(\alpha, \omega) \\
\text{subject to} & \quad Q(\alpha, \omega) \leq 0, \alpha[i] \in \mathbb{N}, \omega[i] \in \mathbb{C},
\end{align*}$$

where $f(\cdot, \cdot)$ is the objective function, $\alpha[i]$ and $\omega[i]$ are the elements of $\alpha$ and $\omega$, and $Q(\cdot, \cdot)$ represents constraints.

Traditional methods for the MINLP problems are often based on mathematical optimization techniques, such as the branch-and-bound (B&B) algorithm that is with high complexity for real-time implementation. An optimal pruning policy has been proposed in [39] to deal with the resource allocation problems in cloud radio access networks (RANs) while a similar method has been utilized in [40] for resource allocation in D2D systems. The resource allocation in [39] and [40] can be formulated as MINLP problems, which can be solved by the globally optimal B&B algorithm. By observation, branching process is most time consuming in the B&B algorithm and a good pruning policy can significantly reduce the computational complexity. The more nodes that would not lead to the optimal solution are pruned, the less time is consumed. Therefore, algorithm acceleration can be formulated into a pruning policy learning task. With invariant problem-independent features and appropriate problem-dependent feature selection, the pruning policy learning task can be further converted into a binary classification problem that can be solved by deep learning techniques. Simulation results of the optimization guided deep learning method for D2D networks are summarized in Table III. In the table, ogap, or optimality gap, means the gap between the optimal solution and the one achieved by the accelerated algorithm, while speed refers to the speedup with respect to the original B&B algorithm. From the table, the accelerated B&B algorithm can achieve good optimality and reduce computational complexity at the same time with only hundreds of training samples.

Besides reducing the sample complexity of deep learning, incorporating prior information of the optimization problem can help handle stochastic constraints of the problem. In many wireless resource allocation problems, the constraints are taken on long term average instead of the instantaneous system performance. The training of DNNs can be undertaken in the dual domain, where the constraints are linearly combined to create a weighted objective. A primal-dual descent approach has been proposed in [56] as a model-free learning approach, where the gradients are estimated by sampling the model functions and wireless channels. From simulation in both the additive Gaussian white noise channel and the interference channel, the performance is close to the well-behaved WMMSE approach [57].

IV. Deep Reinforcement Learning Based Resource Allocation

Deep RL has been found effective in network slicing [58], integrated design of caching, computing, and communication for software-defined and virtualized vehicular networks [59], multi-tenant cross-slice resource orchestration in cellular

| Number of training samples | 90   | 100  | 150  | 200  |
|-----------------------------|------|------|------|------|
| Ogap                        | 3.88%| 3.23%| 2.27%| 2.01%|
| Speed                       | 2.50x| 2.21x| 2.17x| 2.06x|
including both large- and small-scale fading components. Then the received SINR of link $i$ is given by:

$$\gamma_i(t) = \frac{P_i g_{i,i}(t)}{\sum_{j \neq i} P_j g_{j,i}(t) + \sigma^2}, \quad (4)$$

where $P = [P_1, \ldots, P_N]^T$ is the transmit power vector for the $N$ links and $\sigma^2$ represents noise power. Then a dynamic power allocation problem to optimize a generic weighted sum rate is formulated as:

$$\max_{P} \sum_{i=1}^{N} w_i(t) \cdot \log \left( 1 + \gamma_i(t)(P) \right) \quad (5)$$

subject to $0 \leq P_i \leq P_{\text{max}}$, $i \in N$, where $P_{\text{max}}$ is the maximum transmit power for all links and $w_i(t)$ is the nonnegative weight of link $i$ in time slot $t$ that can be adjusted to consider sum rate maximization or proportionally fair scheduling.

Due to the coupling of transmit power across all links, the optimization problem in (5) is in general non-convex and NP-hard [8], not to mention the heavy signaling overhead to acquire global CSI. To address the challenge, a model-free deep RL based power allocation scheme has been developed in [2] that can track the channel evolution and execute in a distributed manner with limited information exchange.

The proposed deep RL method in [21] assumes a centralized training architecture, where each transmitter acts as a learning agent that explores the unknown environment with an $\epsilon$-greedy policy and then sends its exploration experiences to a central controller through backhaul links with some delay. A DQN is trained at the controller using experiences collected from all agents with experience replay, which stores an approximation of action values in different environment states. After training for a while, the updated DQN parameters are broadcast to all agents that use the parameters to construct/update their own DQNs for distributed execution. To have a better representation of the communication environment, the state observed by each agent is constructed to include useful local information (such as its own transmit power in the previous time slot, total interference power and its own channel quality), interference from close interfering neighbors, and generated interference toward impacted neighbors. With a proper reward design that targets system-wide performance, the proposed deep RL method learns to adapt the transmit power of each link only using experiences obtained from interaction with the environment. It has been shown to outperform state-of-the-art, including the WMMSE algorithm in [57] and fractional programming based algorithm in [64], which are assumed to have accurate global CSI. Remarkably, the developed power allocation method leveraging deep RL returns solutions better and faster without assuming prior knowledge about the channels, and thus can handle more complicated but practical nonidealities of real systems. Another interesting observation is that the proposed learning based approach shows impressive robustness in the sense that DQNs trained in a different setting (different initialization or numbers of links) can still achieve decent performance. It suggests using a DQN trained in a simulator to jump-start a newly added node in real communication network is a promising approach. An extension has been developed in [22] that considers each transmitter serves more than one receiver and the data-driven deep RL based approach has been demonstrated to achieve better performance than benchmark algorithms.

The power allocation problem has also been studied in [23, 65], which consider a cellular network with several cells and the base station in each cell serves multiple users. All base stations share the same frequency spectrum, which is further divided into orthogonal sub-bands within the cell for each user, i.e., there exists inter-cell interference but no intra-cell interference. Each base station is controlled by a deep RL agent that takes actions (performs power adaptation) based on its local observation of the environment, including cell power, average reference signal received power, average interference, and a local cell reward. The power control action is discretized to incremental changes of $\{0, \pm 1, \pm 3\}$ dB and the reward is designed to reflect system-wide utility to avoid selfish decisions. The agents take turns to explore the environment to minimize impact on each other that stabilizes the learning process. Each deep RL agent trains a local Q-network using experiences accumulated from its interaction with the communication environment. The deep RL based approach learns to control transmit power for each base station that achieves significant energy savings and fairness among users in the system.

### B. Dynamic Spectrum Access

In its simplest form, the dynamic spectrum access problem considers a single user chooses one of $N$ channels for data transmission, as shown in Fig. 5. In [24], it is assumed that each channel has either “good” or “bad” conditions in each time slot and the channel conditions vary as time evolves. The transmission is successful if the chosen channel is good and unsuccessful otherwise. The user keeps transmitting at successive time slots with the objective of achieving as many successful transmission slots as possible.
The proposed learning based dynamic spectrum access method developed in [24] encounters difficulty when the number of considered scenario [67], [68]. Known), which is known to be optimal or near-optimal in the genie-aided Myopic policy (with channel evolution dynamics defined. But two DNNs are now constructed at the agent, one deep RL framework using Actor-Critic [41] has been further proposed. The learning agent employs the \( \epsilon \)-greedy policy to explore the unknown environment and the accumulated experiences are used to improve the policy with the experience replay technique to break data correlation and stabilize training. When the tested \( N = 16 \) channels are strongly correlated, the proposed learning based dynamic spectrum access method outperforms the implementation friendly model-based Whittle Index heuristic introduced in [66]. It closely approaches the genie-aided Myopic policy (with channel evolution dynamics known), which is known to be optimal or near-optimal in the considered scenario [67], [68].

However, as pointed out in [69], the DQN enabled approach developed in [24] encounters difficulty when the number of channels, \( N \), scales large. To tackle the challenge, a model-free deep RL model is also to maximize the number of successful transmission. Similar to [24], the wireless device constructs a DQN that takes the history of previous actions (TRANSMIT or WAIT) and observations (SUCCESS, COLLISION, or IDLENESS) up to \( M \) time slots, as the input, and the output is the value of each action given the current state. The reward is set to 1 if the observation is SUCCESS and 0 otherwise. The learning agent interacts with the unknown wireless environment to gain experiences for Q-network training without prior knowledge about the protocols that other devices follow. With such a model-free deep RL approach that is purely data-driven, the performance is measurably close to the theoretical upper bound that has full model knowledge.

So far, we have demonstrated the power of deep RL in a single user spectrum access problem. In fact, similar observations translate to the multi-user setting as well, albeit with some variations. As investigated in [71], [72], a set of \( K = \{1, \cdots, K\} \) users attempt transmission over \( N = \{1, \cdots, N\} \) orthogonal channels. In each time slot, every user selects a channel to send its data and the transmission is successful (with an ACK signal received) if no others use the same channel. Different from a single user setting, various design objectives can be defined for multiple users, such as sum rate maximization, sum log-rate maximization (known as proportional fairness), etc., depending on the network utility of interest. Apart from its strong combinatorial nature, the problem is difficult in that the environment is only partially observable to each user and nonstationary from user’s perspective due to the interaction among multiple users when they are actively exploring and learning.

The deep RL based framework developed in [71], [72] assumes a centralized training and distributed implementation architecture, where a central trainer collects the experiences from each user, trains a DQN, and sends the trained parameters to all users to update their local Q-network in the training phase. In the implementation phase, each user inputs local observations into its DQN and then acts according to the network output without any online coordination or information exchange among them. To address partial observability, a long short-term memory (LSTM) layer is added to the DQN that maintains an internal state and accumulates observations over time. The local observation, \( S_i(t) \), of user \( i \) at time slot \( t \) includes its action (selected channel), the selected channel capacity, and received ACK signal in time slot \( t - 1 \). Such observation is implicitly aggregated in the LSTM layer embedded in the DQN to form a history of the agent. The action, \( a_i(t) \), of user \( i \) in time slot \( t \) is drawn according to the
distribution to balance exploitation and exploration

$$\Pr(a_t = a) = \frac{(1 - \alpha)e^{\beta Q(a, S_t(t))}}{\sum_{a \in A} e^{\beta Q(a, S_t(t))}} + \frac{\alpha}{N + 1},$$

(6)

where $\alpha > 0$, $\beta$ is the temperature to be tuned in training, and $Q(a, S_t(t))$ is the value of selecting channel $a$ for a given observation $S_t(t)$ according to the DQN output.

To address the issue of environment nonstationarity, experience replay, which has been popular in most deep RL training but could continuously confuse the agent with outdated experiences in a nonstationary environment, is disabled during training. More advanced techniques, such as dueling network architecture [73] and double Q-learning [74], are leveraged to improve training convergence. When compared with the classical slotted ALOHA protocol, opportunistic channel aware algorithm [75], [76], and distributed protocol developed in [77], the proposed deep RL method consistently achieves better performance in terms of average channel utilization, average throughput, and proportional fairness with properly designed reward according to the utility of interest.

C. Joint Spectrum and Power Allocation: Application Example in Vehicular Networks

Fig. 6. An illustration of spectrum sharing in vehicular networks, where V2V and V2I links are indexed by $k$ and $m$, respectively and each V2I link is preassigned an orthogonal RB.

In a vehicular network as illustrated in Fig. 6 $K$ vehicle-to-vehicle (V2V) links share the spectrum of $M$ vehicle-to-infrastructure (V2I) links to improve its utilization efficiency. The V2I links are designed for high data rate entertainment services while the V2V links need to support reliable dissemination of safety-critical messages, formally stated as the successful delivery of packets of size $B$ within the time constraint $T_0$. Such a reliability requirement of V2V links is hard to handle with traditional optimization approaches due to its exponential complexity in the length of $T_0$. However, it has been shown in [16], [17] that we can nicely treat the issue in the deep RL framework through properly designing a reward that correlates with the objective. We assume that each V2I link has been assigned an orthogonal resource blocks (RBs) and uses a fixed transmit power. Then, each V2V transmitter needs to carefully select the V2I RB to share and adjust its transmit power to avoid strong interference and ensure both V2I and V2V links to achieve their respective goals.

1) Single-Agent RL: In view of the difficulty to collect global CSI at a central controller in real time, a distributed resource allocation algorithm has been developed in [16] that leverages deep RL. In particular, each V2V transmitter serves as a learning agent that occupies a local copy of a DQN and follows the $\epsilon$-greedy policy to explore the unknown environment. The observation of each V2V agent represents its own perception of the unknown environment state, given by

$$S_t = \{G_t, H_t, I_{t-1}, N_{t-1}, L_t, U_t\},$$

(7)

where $G_t$ and $H_t$ represent the current V2V signal channel strength and the interference channel strength from the V2V transmitter to the base station over all RBs, respectively, $I_{t-1}$ is the received interference power, $N_{t-1}$ is the selected RBs of neighbors in the previous time slot over all RBs, $L_t$ and $U_t$ denote the remaining load and time budget to meet latency constraint from the current time slot, respectively. The action of each V2V agent amounts to a selection of RB as well as discrete transmit power levels. The reward balances V2I and V2V requirements, given by

$$r_t = \lambda_c \sum_m C^c[m] + \lambda_v \sum_k C^v[k] - \lambda_p (T_0 - U_t),$$

(8)

where $C^c[m]$ and $C^v[k]$ represent the capacity of V2I link $m$ and V2V link $k$, respectively. $\lambda_c$, $\lambda_v$, and $\lambda_p$ are nonnegative weights to balance different design objectives. In particular, the inclusion of $T_0 - U_t$ in the reward constantly reminds agents of the upcoming deadline for V2V payload transmission and effective helps improve payload delivery rates for V2V links.

The reward design in (8) facilitates system-wide performance improvement. But it also dictates the use of a centralized training architecture, where a central controller collects experiences from all V2V agents and compiles the reward for the DQN training. The system architecture is illustrated in Fig. 7 In the implementation, each V2V agent performs a local observation of the environment and then use its local copy of the trained DQN to guide its RB selection and power control in a distributed manner. To alleviate the impact of environment nonstationarity due to mutual interaction among multiple V2V links, the V2V agent takes turns to change its action to stabilize training, rather than acting simultaneously. Experimental results demonstrate that the proposed deep RL based approach can learn from scratch to perform intelligent spectrum and power allocation that outperforms the benchmarks, including the distributed algorithm in [78] and a random baseline, in terms of both sum V2I rate as well as the delivery rate of V2V payloads.
2) Multi-Agent RL: To further improve the network performance and better handle the dynamics in a vehicular network, we investigate in [17] a multi-agent RL based approach to enable all V2V agents to perform resource allocation simultaneously. We have iterated throughout the article that such simultaneous actions of all learning agents tend to make the environment observed by each agent highly nonstationary and compromissies stability of DQN training. To address the issue, either the experience replay technique that is central to the success of deep RL is disabled as in [71], [72], or each agent take turns to update its action as in [16], [21], [23]. We believe that such a turn-taking action update constraint leads to inevitable sub-optimality as it is a subset of the simultaneous action space and the disabling of highly efficient experience replay techniques is undesirable. In response, we leverage the fingerprint based method proposed in [79] that identifies and addresses the source of nonstationarity: the policy change of other agents due to learning. As such, the environment observed by each agent can be made stationary by conditioning on other agents’ policy change, i.e., we augment the observation of each V2V agent with an estimate of the policy change of all other agents, the idea of hyper Q-learning [80]. Further analysis reveals that the agents’ policy varies along the learning process, whose trajectory can be tracked by a low-dimensional fingerprint, including the training iteration number $e$ and the probability of selecting a random action, $\epsilon$, in the $\epsilon$-greedy policy. Then we revise the observation of each V2V agent as

$$Z(t) = \{S_t, e, \epsilon\}, \quad (9)$$

where $S_t$ contains similar local observations (measurements) of the $k$th V2V agent as in (7). In addition, we revise the reward design by considering the V2V related reward component at each time step to the sum V2V rate when payload delivery is not finished and to a constant number $\beta$ larger than the largest sum V2V rate when finished. The design of $\beta$ reflects the tradeoff between designing purely toward the ultimate goal and learning efficiency. For pure goal-directed consideration, we set the V2V related reward to 0 at each step until the payload is delivered when the reward is 1. However, receiving such delayed rewards decreases training efficiency and hence we impart our domain knowledge by incorporating the sum V2V rate as aggregated rewards. Again we employ the centralized training and distributed implementation architecture in Fig. 7 to train multiple DQNs for V2V agents with experience replay and then deploy them after training is done.

Remarkably, such a simple approach proves very effective in stabilizing DQN training and combined with the revised reward design, it significantly outperforms both the single-agent RL based approach in [16] and a random baseline. In fact, the proposed multi-agent RL based method encourages V2V cooperation despite the distributed implementation without online coordination. To illustrate, we break down in Fig. 8 the change of V2V transmission rates of the proposed method and random baseline over the time constraint $T_0 = 100$ ms for one episode where all V2V links successfully deliver payload using the proposed method. But for the random baseline, Link 3 fails the task and all others succeed. Comparing Figs. 8 (a) and (b), one can tell that the proposed method enables Link
network architectures have varied strengths and weaknesses. Likewise, wireless resource allocation has its own unique characteristics that are worth considering when we adapt or redesign an appropriate DNN. As an example, in power allocation, the adjustment of one transmitter’s power affects not only its intended receiver but other closely located wireless links, making the problem non-convex. However, if we convert it to the dual domain, solving the problem indeed becomes much easier and the duality gap is shown to be small in most practical cases [3]. Such domain knowledge is expected to be very helpful for guiding DNN design in the wireless context. But it is not clear what is the best way to use. In the unsupervised approaches of using deep learning for optimization mentioned earlier, we treat the resource allocation objective as the loss function for learning. More often than not, the objective is different from what we normally use in DNN training, like mean-squared error or cross-entropy. It is unclear whether the existing network architecture best suits the need of resource allocation objective minimization or maximization. Further, theoretical understanding behind the convergence of training DNNs with the new objectives and efficient training techniques largely remain unknown.

B. Bridging the Gap between Training and Implementation

Most, if not all, proposed learning algorithms in the existing literature are trained and tested in an offline simulator. We understand this makes algorithm design and testing quick and easy, and the policy learned from simulated experiences can be used as a jump-starter for real-time deployment, following the idea of transfer learning. Nonetheless, it is extremely difficult to build a simulator with enough fidelity that can make the learned policy work as expected when interacting with the real-world environment. It seems that the only way we can provide an ultimate answer to the puzzle is to perform policy learning in the real-world. The issue is then how to avoid catastrophic situations while the agents are actively exploring the environment and no concrete policies have been obtained yet. A seemingly good answer might be to use expert human knowledge to help confine the exploration space and guide the learning agent’s search within the space. But exactly how to implement the concept in algorithm design with performance guarantee is unclear and worth further research efforts.

C. Multi-Agent Consideration in Deep RL

In the wireless domain, most scenarios that we are interested in are of multi-user nature, whether it is the power control for multiple users within a cell, or joint spectrum and power allocation for multiple V2V links in a vehicular network. In these cases, actions of one user impacts the performance of not only itself but also others nearby. From the perspective of each user, the environment that it observes then exhibits nonstationarity when other users are actively exploring the state and action space for policy learning. Meanwhile, each user can only obtain local observation or measurements of the true underlying environment state, which, in the deep RL terminology, is a partially observable MDP. Then the learning agent needs to construct its belief state from the previous actions and local observations to estimate the true environment state, which is a challenging issue. A partial solution to the problem might be to enable inter-agent communications to encourage multi-user coordination and increase local awareness of the global environment state. That said, such a combination of environment nonstationarity and partial observability makes learning extremely difficult, which is made even worse if we scale the number of users large as in the upcoming internet of things era. More investigation in this direction is thus highly desired.

VI. Conclusion

In this article, we have provided an overview of applying the burgeoning deep learning technology to address wireless
resource allocation with application to vehicular networks. We have discussed the limitations of traditional optimization based approaches and the potentials of deep learning paradigms in wireless resource allocation. In particular, we have described in detail how to leverage deep learning to help solve difficult optimization problems for resource allocation and deep RL for a direct answer to many resource allocation problems that cannot be handled or even modeled in the optimization framework. We have further identified some open issues and research directions that warrant future investigation.

References

[1] Z. Han and K. J. R. Liu, Resource allocation for wireless networks: Basics, techniques, and applications. Cambridge university press, 2008.
[2] C. E. Shannon, “A mathematical theory of communication,” Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, Jul. 1948.
[3] W. Yu and R. Lai, “Dual methods for nonconvex spectrum optimization of multicarrier systems,” IEEE Trans. Commun., vol. 54, no. 7, pp. 1310–1322, Jul. 2006.
[4] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, “Multiuser OFDM with adaptive subcarrier, bit, and power allocation,” IEEE J. Sel. Areas Commun., vol. 17, no. 10, pp. 1747–1758, Oct. 1999.
[5] G. Song and Y. Li, “Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks,” IEEE Commun. Mag., vol. 43, no. 12, pp. 127–134, Dec. 2005.
[6] ———, “Cross-layer optimization for OFDM wireless networks—part I: Theoretical framework,” IEEE Trans. Wireless Commun., vol. 4, no. 2, pp. 614–624, 2005.
[7] ———, “Cross-layer optimization for OFDM wireless networks—part II: algorithm development,” IEEE Trans. Wireless Commun., vol. 4, no. 2, pp. 625–634, 2005.
[8] Z.-Q. Luo and S. Zhang, “Dynamic spectrum management: Complexity and duality,” IEEE J. Sel. Topics Signal Process., vol. 2, no. 1, pp. 57–73, Feb. 2008.
[9] M. Bennis, M. Debbah, and H. V. Poor, “Ultra reliable and low-latency wireless communication: Tail, risk, and scale,” Proc. IEEE, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.
[10] T. O’Shea and J. Hoydis, “An introduction to deep learning for the physical layer,” IEEE Trans. Cogn. Commun. Netw., vol. 3, no. 4, pp. 563–575, Dec. 2017.
[11] S. Dmer, S. Cammerer, J. Hoydis, and S. ten. Brink, “Deep learning based communication over the air,” IEEE J. Sel. Areas Commun., vol. 12, no. 1, pp. 132–143, Feb. 2018.
[12] N. Samuel, T. Diskin, and A. Wiesel, “Learning to detect,” IEEE Trans. Signal Process., vol. 67, no. 10, pp. 2554–2564, May 2019.
[13] N. Farsad and A. Goldsmith, “Neural network detection of data sequences in communication systems,” IEEE Trans. Signal Process., vol. 66, no. 21, pp. 5663–5678, Nov. 2018.
[14] H. Ye, G. Y. Li, and B. Jiang, “Power of deep learning for channel estimation and signal detection in OFDM systems,” IEEE Wireless Commun. Lett., vol. 7, no. 1, pp. 114–117, Feb. 2018.
[15] H. Ye, L. Liang, G. Y. Li, and B.-H. F. Juang, “Deep learning based end-to-end wireless communication systems with conditional GAN as unknown channel,” arXiv preprint arXiv:1903.02531, 2019.
[16] H. Ye, G. Y. Li, and B.-H. Juang, “Deep reinforcement learning based resource allocation for V2V communications,” IEEE Trans. Veh. Technol., vol. 68, no. 4, pp. 3163 – 3173, Apr. 2019.
[17] L. Liang, H. Ye, and G. Y. Li, “Spectrum sharing in vehicular networks based on multi-agent reinforcement learning,” to appear in IEEE J. Sel. Areas Commun., Sep. 2019.
[18] C. Zhang, P. Patras, and H. Haddadi, “Deep learning in mobile and wireless networking: A survey,” to appear in IEEE Commun. Surveys Tut., pp. 1–19, 2019.
[19] Z. Chang, L. Lei, Z. Zhou, S. Mao, and T. Ristaniemi, “Learn to cache: Machine learning for network edge caching in the big data era,” IEEE Wireless Commun., vol. 25, no. 3, pp. 28–35, Jun. 2018.
[20] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, “Wireless network intelligence at the edge,” arXiv preprint arXiv:1812.02858, 2018.
[21] Y. S. Nasir and D. Guo, “Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks,” to appear in IEEE J. Sel. Areas Commun., Sep. 2019.
O. Naparstek and K. Cohen, “Deep multi-user reinforcement learning for dynamic spectrum access,” IEEE Trans. Wireless Commun., vol. 18, no. 1, pp. 310–323, Jan. 2019.

Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas, “Dueling network architectures for deep reinforcement learning,” arXiv preprint arXiv:1511.06561, 2015.

H. Van Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double q-learning,” in Proc. AAAI, 2016, pp. 2094–2100.

K. Cohen and A. Leshem, “Distributed game-theoretic optimization and management of multichannel ALOHA networks,” IEEE/ACM Trans. Netw., vol. 24, no. 3, pp. 1718–1731, Jun. 2016.

T. To and J. Choi, “On exploiting idle channels in opportunistic multichannel ALOHA,” IEEE Commun. Lett., vol. 14, no. 1, pp. 51–53, Jan. 2010.

I.-H. Hou and P. Gupta, “Proportionally fair distributed resource allocation in multiband wireless systems,” IEEE/ACM Trans. Netw., vol. 22, no. 6, p. 18191830, Jun. 2014.

M. I. Ashraf, M. Bennis, C. Perfecto, and W. Saad, “Dynamic proximity-aware resource allocation in vehicle-to-vehicle (V2V) communications,” in Proc. IEEE GLOBECOM Workshops. IEEE, 2016, pp. 1–6.

J. Foerster, N. Dardelli, G. Farquhar, T. Afouras, P. H. S. T. 1, P. Kohli, and S. Whiteson, “Stabilising experience replay for deep multi-agent reinforcement learning,” in Proc. Int. Conf. Mach. Learning (ICML), 2017, pp. 1146–1155.

G. Tesauri, “Extending Q-learning to general adaptive multi-agent systems,” in Proc. Advances Neural Inf. Process. Syst. (NeurIPS), 2004, pp. 871–878.

L. Wang, H. Ye, L. Liang, and G. Y. Li, “Learn to allocate resources in V2X networks,” submitted to Advances Neural Inf. Process. Syst. (NeurIPS), 2019.