Deep Reinforcement Learning Based Relaying for Buffer-Aided Cooperative Communications

Chadi Abou-Rjeily, *Senior Member, IEEE*, and Sawsan El-Zahr

Abstract

The advances in deep reinforcement learning (DRL) have shown a great potential in solving physical layer-related communication problems. This paper investigates DRL for the relay selection in buffer-aided (BA) cooperative networks. The capability of DRL in handling highly-dimensional problems with large state and action spaces paves the way for exploring additional degrees-of-freedom by relaxing the restrictive assumptions around which conventional cooperative networks are usually designed. This direction is examined in our work by advising and analyzing advanced DRL-based BA relaying strategies that can cope with a variety of setups in multifaceted cooperative networks. In particular, we advise novel BA relaying strategies for both parallel-relaying and serial-relaying systems. For parallel-relaying systems, we investigate the added value of merging packets at the relays and of activating the inter-relay links. For serial-relaying (multi-hop) systems, we explore the improvements that can be reaped by merging packets and by allowing for the simultaneous activation of sufficiently-spaced hops. Simulation results demonstrate the capability of DRL-based BA relaying in achieving substantial improvements in the network throughput while the adequate design of the reward/punishment in the learning process ensures fast convergence speeds.

Index Terms

Cooperative Networks, Relaying, Buffers, Reinforcement Learning, 5G Networks.

I. Introduction

A. Relaying and 5G Communications

In cooperative wireless communication networks, nodes can share their resources and act as relays for assisting the communications between other nodes. Relaying networks can cope with

The authors are with the Department of Electrical and Computer Engineering of the Lebanese American University (LAU), PO box 36 Byblos 961, Lebanon. (e-mails: chadi.abourjeily@lau.edu.lb and sawsan.elzahr@lau.edu).

diverse radio propagation conditions and are useful for coverage and capacity improvements. 3GPP LTE-Advanced and IEEE WiMAX both include relaying as one of their key features incorporating in-band/out-band relaying as well as transparent/non-transparent relay connectivity with users [1]. Supporting the increasing demand for data usage and wireless connectivity, the 5G network architecture witnessed a categorical shift from the base station (BS) centric to the user centric paradigm where the wireless nodes are envisaged to participate in storage, relaying and computation within the network as shown in Fig. 1. Moving from structured cellular networks to more unstructured forms of heterogeneous networks (HetNets) [2], 5G relaying must support a multitude of devices and applications with diversified requirements in terms of latency, throughput and reliability [3].

The 5G emerging applications include Device-to-Device (D2D) communications where the coordinated communication with the BS can be bypassed. In this context, devices can communicate either directly or through other devices present in the proximity. In D2D communication setups, relay nodes not only relay information between the BS and user equipment (UE), but also between different UEs for sharing relevant contents [4]. Relaying is also popular with Machine-to-Machine (M2M) communications for the Internet-of-Things (IoT) applications that involve automated data generation, processing and transfer. In IoT networks, some machines might have no information to transmit at certain times. These idle machines can be used as relays to support the communications between the active nodes in the high-density network. For such setups, relay selection is pivotal for increasing the network coverage and enabling the efficient data transfer in the IoT [5].

Relay-assisted communication is pivotal for unmanned aerial vehicles (UAVs) that can be rapidly deployed to support emergency communications in case of disasters or supplement the overloaded existing ground network infrastructure [6]. Different forms of relaying can be envisaged with UAVs. These include single UAV relaying networks where a source and destination located on the ground communicate through a UAV when direct communications are not possible because of the excessive distance and/or presence of obstacles. In this case, the UAV can dynamically adjust its position according to the changes in the environment to achieve the best communication quality. Multi-hop UAV relaying is another form of UAV relay-assisted communications where the communication between the ground nodes is realized through a cascade of UAVs that advantageously communicate with each other over reliable, shorter and unobstructed aerial communication links.

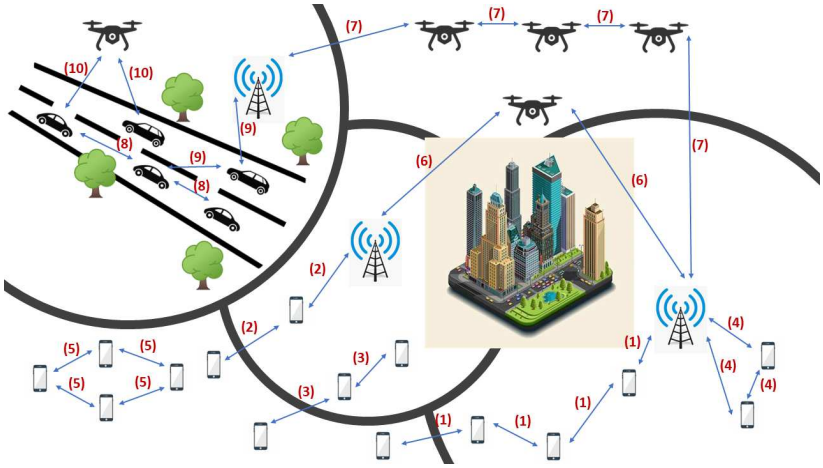

Fig. 1. Relaying in 5G networks. (1): Multi-hop communications. (2): BS-to-Device coverage extension. (3): Device-to-Device coverage extension. (4): Capacity enhancement. (5): Out-of-coverage D2D relaying. (6): Single UAV relaying. (7): Multi-hop UAV relaying. (8): Vehicle-to-Vehicle relaying. (9): Vehicle-to-Infrastructure relaying. (10): UAV assisting communications between two vehicles.

With the commercialization of 5G technology, the Internet-of-Vehicles (IoV) is constantly maturing [7]. IoV is characterized by strong mobility and large amount of information exchange rendering the conventional point-to-point communications incapable of meeting the Quality-of-Service (QoS) demands in such complex and changeable communication environments. In IoV, implementing efficient relay selection strategies is crucial for improving the spectral efficiency and reducing the delays. Such strategies must take into consideration the real-time changes in the dynamic network that operates under strong interference conditions.

B. Cooperative Network Architecture and Relaying Methods

The relaying protocols can be categorized into two principal classes: Amplify-and-Forward (AF) and Decode-and-Forward (DF). AF is a nonregenerative solution where the signal received at the relay is simply amplified before being retransmitted unlike the regenerative DF strategy where the received signal is decoded prior to retransmission. The common architectures of the cooperative networks that were widely investigated in the open literature are parallel and serial relaying. In parallel-relaying, the relay nodes receive the message broadcasted by the source node in one time slot and cooperate with each other to deliver this message to the destination node in the subsequent slot. Since the message reaches the destination along a multitude of paths that are subject to different fading conditions, parallel-relaying networks achieve spatial diversity in a

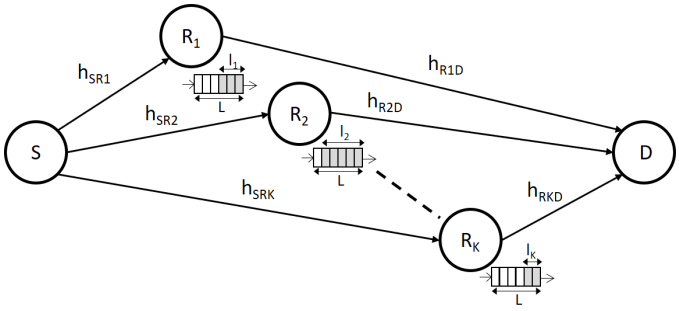

Fig. 2. BA parallel-relaying with K relays in the vicinity of the source (S) and destination (D). Each relay is equipped with a buffer of size L.

distributed manner. In particular, a diversity order equal to the number of relays can be achieved, thus, enhancing the reliability of the communications. Serial (or multi-hop) relaying refers to scenario where the information is sent from a source to a destination via a set of relays in cascade. This technique is beneficial for extending the network coverage in case the terminal nodes are separated by long distances that render the direct transmissions inefficient. Conventionally, relays operate in the half-duplex (HD) mode and are restricted to receive and transmit over orthogonal channels (in frequency or time). Recently, the interest in full-duplex (FD) relaying is on the rise as a means of improving the spectral efficiency by deploying relays that can support the concurrent reception and transmission in the same frequency band. However, FD relaying is advantageous only if powerful self-interference cancellation techniques are implemented at the relays to suppress the residual interference from the transmitting to the receiving circuits [8]. A relay misbehaviour detection scheme was proposed in [9] where some relays might not operate in a normal or trustworthy manner.

C. Conventional Buffer-Aided (BA) Relaying

The relaying strategies have evolved from being buffer-free (BF) to become buffer-aided (BA) with storage capabilities enabled at the relays. In cooperative communication systems, equipping the relays with buffers provides an additional degree-of-freedom to combat fading since the information packets can be temporarily stored until the channel conditions become more favorable [10]. This approach improves the network's reliability and throughput at the expense of introducing queuing delays [11], [12].

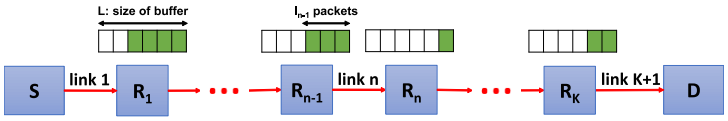

Fig. 3. BA serial-relaying with K relays ($K+1$ hops). Each relay is equipped with a buffer of size L.

Relay selection is often adopted in parallel-relaying networks with a single node being activated in each time slot as shown in Fig. 2. This approach limits the signalling overhead and leverages the synchronization requirements. For BF systems, the *max-min* protocol can be implemented where the information is relayed through the relay with the highest signal-to-noise ratio (SNR). The SNR of an end-to-end link is equal to the minimum of the SNRs along the source-relay and relay-destination hops implying that the weakest of these hops dictates the error performance regardless of the quality of the other hop. As such, for DF relaying with HD relays, the *max-min* protocol achieves a diversity order equal to the number of relays. This constraint of inflicting the reception and transmission through the same relay in two consecutive time slots can be leveraged with BA relaying. In fact, since the relays possess storage capabilities, one relay can be selected for reception in one time slot while a different relay might be selected for transmission in the subsequent slot where the incoming packet is stored at the receiving relay's buffer while the retransmitted packet is extracted from the buffer of the transmitting relay. For example, *max-link* relaying is a suitable method for realizing BA cooperation in DF-HD networks [13]. For this scheme, the strongest available link, among all source-relay and relay-destination links, is activated resulting in the maximum diversity order that is equal to twice the number of relays. In this context, a relay might be selected for reception or transmission if its buffer is not full or empty, respectively. A comparable BA parallel-relaying protocol was advised in [14] where the priority was given to the source-relay and relay-destination hops in odd and even time slots, respectively. This scheme resulted in slight improvements in the diversity order compared to the *max-link* protocol with finite-size buffers.

While the relay selection in [13], [14] is based solely on the channel state information (CSI), more recent BA relaying protocols include the buffer state information (BSI) in the relay selection process as well [15]–[19]. The rationale is to include both the quality of the links and the number of stored packets in the relay selection mechanism. This embraced policy incentivizes the transmission from congested buffers and the reception at under-filled buffers thus ensuring

a smooth flow of packets from the source to the destination. Balancing the loads of the relays' buffers limits the queuing delays and guarantees improved performance levels with practical buffers that have small sizes in contrast to the *max-link* protocol that is beneficial with infinite-size buffers. For example, the work in [19] developed a threshold-based relay selection scheme where K threshold levels were fixed at the K relays. Based on the difference between the actual number of stored packets and the threshold level, different classes of priority were assigned to the relays. As such, by adjusting the threshold levels, different levels of tradeoff between the outage probability (OP) and average packet delay (APD) were achieved. In particular, the diversity order of $K+N$ and the asymptotic APD of $2N+2$ can be achieved where the integer N depends on the K threshold levels and ranges from 0 and K.

BA relaying is also appealing for multi-hop communications as shown in Fig. 3. In fact, the performance of a BF serial-relaying network is dominated by the weakest of its hops. Despite the realised coverage extension, the achievable diversity order with multi-hop BF relaying is equal to one underscoring no diversity gains compared to the conventional non-cooperative point-to-point communications. This limitation was leveraged by instigating BA relaying. CSI-based BA serial-relaying was studied in [20] where the available hop with the highest instantaneous SNR is selected. This type of selection results in a diversity order that is equal to the number of hops when the deployed buffers have an infinite size. CSI-BSI-based BA serial-relaying was suggested in [21] where buffer-occupancy-related weights were assigned to the hops that are not in outage and the hop with the highest weight was selected. Through a Markov chain analysis, this methodology was proven to achieve the full diversity order (equal to the number of hops) with finite-size buffers while benefiting from bounded delays that increase with the number of hops but not with the buffer size.

D. DRL-based Buffer-Aided (BA) Relaying

Reinforcement Learning (RL) is an important branch of Machine Learning (ML) that has attracted an increasing interest recently [22]. Unlike supervised deep learning (DL) where the system is trained with a sample dataset to extract some useful features in a highly-dimensional space, RL involves an intelligent agent that interacts with the environment in order to maximize the notion of a cumulative reward. The environment is typically formulated as a Markov decision process (MDP) that comprises a number of states. At a current state, the agent takes an action, receives an immediate reward and moves to a next state. Based on such experiences, the agent

adjusts its policy to achieve the optimal policy. In order to efficiently attain this policy in complicated system models that involve very large state and/or action spaces, deep reinforcement learning (DRL) algorithms are often applied based on combining RL and DL. DRL algorithms take advantage of the powerful function-approximating capabilities of deep neural networks (DNNs) in order to improve the learning speed and tackle high-dimensional complex problems.

Q-learning is a widely used RL algorithm that iteratively updates its value estimates for each state-action pair based on the rewards observed during exploration. Specifically, Q-learning uses the Bellman equation to update its Q-value estimates where this process is repeated until the Q-values converge to their optimal values resulting in a learned policy that maximizes the expected cumulative reward. Deep Q-learning is a more advanced version of Q-learning that uses a deep neural network to estimate the Q-values instead of a lookup table. The evolution process of Deep Q-learning is similar to that of Q-learning, with the added step of updating the parameters of the neural network using backpropagation.

DRL emerged as a powerful tool to effectively address various challenges in the area of communications and to solve physical layer-related problems including network access, adaptive rate control, proactive caching, data/computation offloading, network security and connectivity preservation as well as the detection of abnormal traffic in networks [23]–[25]. DRL can be used for dynamic spectrum access where sensors make independent decisions on the selected channel. For example, in order to maximize the throughput, the RL agent receives a positive reward if the selected channel suffers from low interference and gets a negative reward otherwise. In IoVs, DRL-based power allocation increases the number of vehicles meeting the latency constraint. In this case, the reward is a function of the user's capacity and latency. DRL is also applied for resource allocation with energy harvesting-enabled IoT devices with the objective of maximizing the IoT network lifetime. Other physical layer-related topics that were handled using DRL include deceiving jammers in wireless networks and the joint user association and channel selection in HetNets.

Beside the aforementioned applications of DRL for the physical layer, DRL recently attracted an increasing interest for advising relay selection strategies in BA cooperative networks. The work in [26] considered the parallel-relaying setup shown in Fig. 2 with DF-HD relays. The target of the relay selection strategy in [26] was to maximize the number of packets delivered to the destination node for a communication session that extends over a number of time slots subject to the two following constraints. (i): The delay of the packet delivered to the destination

must not exceed a certain target packet delay. (ii): At most one packet can be communicated along the source-relay or relay-destination links in each time slot in order to avoid interference and respect the HD constraint. For such setups, a state of the MDP comprises the numbers of packets stored in the relays' buffers as well as the availability of the source-relay and relay-destination communication links. Since the number of states increases exponentially with the number of relays, Q-learning is not appropriate for solving this relay selection problem. In fact, the dimensions of the Q-table that contains the Q-values of all state-action pairs will be huge motivating the implementation of DRL where DNNs are used to evaluate the Q-values. In order to accelerate the convergence and guide the learning mechanism on how to tackle invalid actions, a decision-assisted learning approach was adopted in [26] by including extra training-pairs in the generated experiences. These experiences correspond to invalid actions for which the Q-values in the target network are imposed to be zero. The invalid actions correspond to the transmission from an empty buffer and the reception at a full buffer. The invalid actions also include the activation of a communication link that does not meet the target rate requirement.

DRL-based DF-HD BA parallel-relaying was also considered in [27] where the system comprised two destination nodes and each relay was equipped with two buffers to serve each one of the users. A throughput-maximization approach under a delay constraint was adopted similar to [26]. The optimization problem in [27] also considered switching between the orthogonal multiple access (OMA) and nonorthogonal multiple access (NOMA) transmission modes along with optimizing the NOMA power allocation factor. In [28], DRL was applied for BA relay selection in parallel-relaying cognitive networks in the presence of an eavesdropper. Two optimization problems were formulated and solved using DRL; one maximizes the throughput subject to delay and secrecy constraints and the second one targets maximizing the secrecy rate under the delay constraints. In [28], it was assumed that the relays can operate either in the HD or FD modes. In the former case, a relay can only receive a signal from the secondary source while in the latter case it can simultaneously receive this signal and transmit a jamming signal to interfere with the eavesdropper. For the implementation of DRL in both [27] and [28], a positive reward was given if a packet arrives at the destination with a delay not exceeding the target delay while a negative reward was inflicted to discourage invalid actions. Moreover, in [27], [28], the invalid actions were removed from the action set at the output of the DNN in order to reduce the range of exploration and improve the convergence speed.

E. Contributions

The existing literature on DRL-based DF-HD BA relaying can be further leveraged as follows. (i): The existing schemes [26]–[28] are all fixed-rate schemes where, at most, a single information packet can be transmitted along each link. (ii): The BA parallel-relaying schemes [26]–[28] consider a relatively simple network architectures where the relays can only communicate with the source and destination. In other words, the possibility of activating inter-relay links was overlooked. (iii): To the authors' best knowledge, DRL was limited to parallel-relaying networks and there are no existing works that advise DRL for multi-hop communications.

As such, the contributions of this work are two-fold:

- We propose DRL-based relaying schemes for BA parallel-relaying networks with more sophisticated communication paradigms that can extract the full capabilities of the underlying cooperative network. In particular, we consider the possibility of simultaneously transmitting more than one information packet along high-SNR links and we investigate the advantage of enabling inter-relay communications.
- We propose DRL-based relaying protocols for BA serial-relaying networks. These protocols embed the aforementioned quantized adaptive-rate transmission. Moreover, unlike the parallel-relaying case, the serial-relaying protocols can support the simultaneous transmissions from multiple sufficiently-spaced relays.

The additional degrees-of-freedom in the DF-HD relaying protocols incur a considerable growth in the state and action spaces. As such, the DRL algorithms must be adequately designed to ensure acceptable convergence speeds.

II. System Model

Different setups will be considered for parallel and serial relaying. For all setups, we denote by K the number of relays that are assumed to be HD and operate in the DF mode. The relays will be denoted by R_1, ..., R_K. We also assume that there is no direct path linking S to D because of the large distance separating these nodes, for example. Moreover, each relay is equipped with a buffer of size L and we denote by l_k the number of packets stored in the buffer of the k-th relay R_k for $k = 1, \ldots, K$ where $l_k \in \{0, \ldots, L\}$. Finally, Rayleigh block fading is assumed and all links are corrupted with an additive white Gaussian noise (AWGN). Rayleigh fading constitutes the most general and widely spread distribution often adopted to model wireless channels in the absence of a line-of-sight. This fading model has been adopted in the previous works on BA

relaying systems [11]–[21], [26]. Given the abundance of channel models in the literature, other fading models can be readily applied with the only implication of altering the expressions of the outage probabilities without affecting the core technical contributions of this work. In other words, the proposed DRL framework does not entail any restrictions on the channel model to be used.

We assume that the cooperative network is to operate at a target rate of r_0 (in bits per channel use (BPCU)). As such, a communication link will be in outage if its channel capacity falls below that target rate r_0. In this case, a link that suffers from outage cannot be activated by the relay selection protocol since there is no guarantee that the transmitted packet will reach the receiving node with an arbitrarily small probability of error.

A. Parallel-Relaying

Consider the parallel-relaying setup shown in Fig. 2 where the information packets are relayed from S to D through a cluster of K relays. The practical applications of this setup are delineated in scenarios (4) and (5) shown in Fig. 1. The network comprises $2K$ S-R and R-D links among which, at most, a single link can be activated within each time slot in order to avoid interference. We denote by h_k and h'_k the channel coefficients of the S-R$_k$ and R$_k$-D links, respectively. The coefficient h_k (resp. h'_k) is assumed to be a circularly symmetric complex Gaussian distributed random variable with zero mean and variance Ω_k (resp. Ω'_k).

The channel capacities C_k and C'_k of the S-R$_k$ and R$_k$-D links, respectively, are given by:

$$C_k = \frac{1}{2}\log_2(1+\bar{\gamma}|h_k|^2) \quad ; \quad k=1,\ldots,K \tag{1}$$

$$C'_k = \frac{1}{2}\log_2(1+\bar{\gamma}|h'_k|^2) \quad ; \quad k=1,\ldots,K, \tag{2}$$

where $\bar{\gamma}$ stands for the average signal-to-noise ratio (SNR). In (1)-(2), the division by two follows since a packet transmitted from S needs two hops to reach D.

We consider four setups with parallel-relaying; namely, Single-Packet with No Inter-relay cooperation (SPNI), Multiple-Packets with No Inter-relay cooperation (MPNI), Single-Packet With Inter-relay cooperation (SPWI) and Multiple-Packets With Inter-relay cooperation (MPWI).

1) Single-Packet with No Inter-relay cooperation (SPNI): This constitutes the benchmark scheme considered in [26] where a single S-R or R-D link is activated in each time slot with a single packet transmitted along this link.

The link S-R$_k$ is available if $C_k \geq r_0$ and the buffer at R$_k$ is not full ($l_k \neq L$) so that the incoming packet can be stored. Similarly, the link R$_k$-D is available if $C'_k \geq r_0$ and the buffer at R$_k$ is not empty ($l_k \neq 0$) so that a packet can be extracted and communicated to D. As such, for SPNI, the maximum number of packets per link is given by:

$$n_{max}(k) = \begin{cases} \min\{\delta_{C_k \geq r_0}, L - l_k\}, & \text{S-R}_k \text{ link for } k = 1, \cdots, K; \\ \min\{\delta_{C'_k \geq r_0}, l_k\}, & \text{R}_k\text{-D link for } k = 1, \cdots, K. \end{cases}, \tag{3}$$

where $\delta_{\mathcal{S}} = 1$ if the statement $\mathcal{S}$ is true and $\delta_{\mathcal{S}} = 0$ otherwise.

From (3), it can be observed that $n_{max}(k)$ can be either 0 (no packet is communicated) or 1 (a single packet is communicated). In fact, when $C_k < r_0$ (resp. $C'_k < r_0$), $\delta_{C_k \geq r_0} = 0$ (resp. $\delta_{C'_k \geq r_0} = 0$) implying that $n_{max}(k) = 0$ and no packet can be communicated along the link S-R$_k$ (resp. R$_k$-D) since this link is in outage. Otherwise, when the link S-R$_k$ is not in outage, $\delta_{C_k \geq r_0} = 1$ implying that $n_{max}(k) = 1$ when $l_k < L$ (the buffer is not full) and $n_{max}(k) = 0$ when $l_k = L$ (the buffer is full). Similarly, when the link R$_k$-D is not in outage, $\delta_{C'_k \geq r_0} = 1$ implying that $n_{max}(k) = 1$ when $l_k > 0$ (the buffer is not empty) and $n_{max}(k) = 0$ when $l_k = 0$ (the buffer is empty).

2) Multiple-Packets with No Inter-relay cooperation (MPNI): This scheme allows for transmitting more than one packet along the selected S-R or R-D links. In fact, a S-R$_k$ (resp. R$_k$-D) communication link can support the reliable communication of a number of packets that is equal to $\lfloor C_k/r_0 \rfloor$ (resp. $\lfloor C'_k/r_0 \rfloor$) where $\lfloor \cdot \rfloor$ stands for the flooring operation. This form of quantized adaptive-rate transmission is practically easy to implement by applying high-order modulations at high SNRs for the sake of combining and encoding multiple information packets.

As in (3), the maximum number of packets per link must account for the quality of the link and the buffer state as follows:

$$n_{max}(k) = \begin{cases} \min\{\lfloor C_k/r_0 \rfloor, L - l_k\}, & \text{S-R}_k \text{ link for } k = 1, \cdots, K; \\ \min\{\lfloor C'_k/r_0 \rfloor, l_k\}, & \text{R}_k\text{-D link for } k = 1, \cdots, K. \end{cases}, \tag{4}$$

where, for example, even if the link R$_k$-D can reliably carry $\lfloor C'_k/r_0 \rfloor$ packets but only $l_k < \lfloor C'_k/r_0 \rfloor$ packets are stored in the buffer of R$_k$, then the merged packet can comprise only l_k packets. Similarly, for the link S-R$_k$ (resp. R$_k$-D), $n_{max}(k) = 0$ when $l_k = L$ (resp. $l_k = 0$) where the buffer at R$_k$ is full (resp. empty).

3) Single-Packet With Inter-relay cooperation (SPWI): As with SPNI, at most one packet can be transmitted along each link. However, in addition to the S-R and R-D links, the inter-relay

R-R links can be activated as well where a packet might flow from a relay to a subsequent relay (if any). While (3) captures the maximum number of packets along the S-R and R-D links, the number of packets that can be communicated along the R_k-R_{k+1} link is given by:

$$n_{max}(k) = \min\{\delta_{C_k'' \geq r_0}, l_k, L - l_{k+1}\} \quad ; \quad \mathrm{R}_k\text{-}\mathrm{R}_{k+1} \text{ link } ; \ k = 1, \cdots, K-1, \tag{5}$$

where $C_k'' = \frac{1}{2}\log_2(1 + \bar{\gamma}|h_k''|^2)$ stands for the capacity of the R_k-R_{k+1} link with h_k'' standing for the channel coefficient of this link. From (5), for $n_{max}(k)$ to be different from zero, a necessary condition is that the buffer at the transmitting relay R_k must not be empty ($l_k \neq 0$) and the buffer at the receiving relay R_{k+1} must not be full ($l_{k+1} \neq L$).

4) Mingle-Packets With Inter-relay cooperation (MPWI): For this setup, one link is selected among the available S-R, R-D and R-R links with the possibility of transmitting multiple packets along the selected link. At every time slot, the maximum number of packets that can be transmitted follows from (4) and the following relation

$$n_{max}(k) = \min\{\lfloor C_k''/r_0 \rfloor, l_k, L - l_{k+1}\} \quad ; \quad \mathrm{R}_k\text{-}\mathrm{R}_{k+1} \text{ link } ; \ k = 1, \cdots, K-1, \tag{6}$$

accounting for the possibility of simultaneously transmitting up to $\lfloor C_k''/r_0 \rfloor$ packets if this number of packets is available at the transmitting buffer and if there is enough storage space available at the receiving buffer.

The rationale behind including the R-R links in SPWI and MPWI is as follows. When the S-R and R-D links are unavailable, initiating a communication among two consecutive relays contributes to the flow of packets from congested buffers to under-filled buffers. This accomplished form of load-balancing boosts the availability of the links in the subsequent time slot since relays with full (resp. empty) buffers are unavailable for reception (resp. transmission). This is especially true given the broadcast nature of radio-frequency (RF) transmissions where a message transmitted from a relay can be overheard by D and by the neighboring relays.

B. Serial-Relaying

Consider the serial-relaying setup shown in Fig. 3 where the information transmitted from S flows in tandem from one relay to another until it reaches D. The practical contexts of this setup are provided in scenarios (1), (2), (3) and (7) in Fig. 1, for example. Note that the single-relay setups in scenarios (6), (8), (9) and (10) in Fig. 1 can fall under either the parallel-relaying or serial-relaying classification. For a K-relay system, we denote S and D by R_0 and R_{K+1},

respectively. In this case, the serial-relaying (multi-hop) system comprises $K+1$ hops. Denoting by h_k the channel coefficient along the k-th hop between R_{k-1} and R_k, the channel capacity of this link can be determined from:

$$C_k = \frac{1}{K+1}\log_2(1+\bar{\gamma}|h_k|^2), \tag{7}$$

where the variance of h_k is denoted by Ω_k. Unlike (1)-(2), the division by $K+1$ is introduced in (7) since the communication of a packet from S to D is performed in $K+1$ time slots.

It is worth highlighting that multi-hop relaying is a physical-layer fading mitigation technique that differs substantially from the transport-layer routing problem. In point-to-point communications, information will be lost if the channel is in fading; however, in multi-hop relaying systems the message is transferred from S to D along numerous shorter hops with better propagation conditions resulting in an improved reliability of the end-to-end communications. As such, the information packets flow in a predefined manner from S to R_1, then R_1 to R_2 ... until the packets reach D as shown in Fig. 3. Consequently, the existence of a single end-to-end path from S to D renders the relaying problem different from the routing problem that consists of selecting an appropriate route from several potential routes that might be established in a dense network. In this context, since the physical-layer serial-relaying solution is optimized independently from the other communication layers, then any routing solution at the transport-layer might be applied on top of the proposed relaying strategy.

Four setups will be considered with serial-relaying; namely, Single-Link Single-Packet (SLSP), Single-Link Multiple-Packets (SLMP), Multiple-Links Single-Packet (MLSP) and Multiple-Links Multiple-Packets (MLMP).

1) Single-Link Single-Packet (SLSP): For this scheme, one packet is transmitted along the single selected link as in [20], [21].

The availability of a link is related to the buffer states where the buffer at the transmitting node should contain at least one packet and the buffer at the receiving node must not be full. As such, the number of packets that can be transmitted along the k-th hop is given by:

$$n_{max}(k) = \begin{cases} \min\{\delta_{C_k \geq r_0}, L-l_1\}, & \text{S-R}_1 \text{ link } (k=1); \\ \min\{\delta_{C_k \geq r_0}, l_K\}, & \text{R}_K\text{-D link } (k=K+1); \\ \min\{\delta_{C_k \geq r_0}, l_{k-1}, L-l_k\}, & \text{R}_{k-1}\text{-R}_k \text{ link for } k=2,\cdots,K. \end{cases}, \tag{8}$$

where the source is assumed to be infinitely backlogged implying that a packet is always available at this node.

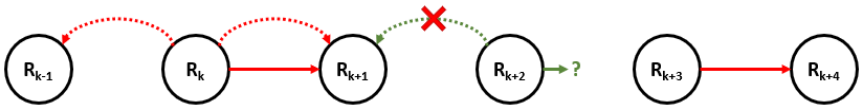

Fig. 4. Interference constraint for the activation of multiple links in serial-relaying systems.

From (8), the case of $l_{k-1} = 0$ means that the buffer at the sending relay is empty and, hence, no packet can be transmitted over the link R_{k-1}-R_k. Similarly, if $l_k = L$, then the number of remaining slots in the receiving relay is $L - l_k = 0$ and, consequently, no additional packets can be accommodated at R_k. Finally, if $C_k < r_0$, then $\delta_{C_k \geq r_0} = 0$ which means that the k-th link is in outage and cannot be activated.

2) Single-Link Multiple-Packets (SLMP): For this scheme, only one hop is activated with the possibility of transmitting more than one packet along this hop. Replacing the factor $\delta_{C_k \geq r_0}$ in (8) (that limited the transmission to one packet) by $\lfloor C_k/r_0 \rfloor$ results in:

$$n_{max}(k) = \begin{cases} \min\{\lfloor C_k/r_0 \rfloor, L - l_1\}, & \text{S-R}_1 \text{ link } (k = 1); \\ \min\{\lfloor C_k/r_0 \rfloor, l_K\}, & \text{R}_K\text{-D link } (k = K+1); \\ \min\{\lfloor C_k/r_0 \rfloor, l_{k-1}, L - l_k\}, & \text{R}_{k-1}\text{-R}_k \text{ link for } k = 2, \cdots, K. \end{cases} . \tag{9}$$

3) Multiple-Links Single-Packet (MLSP): For this setup, the interference constraint is loosened and more than one non-interfering hop can be activated without packet merging. It is assumed that when R_k transmits a packet, it interferes only with the previous relay R_{k-1} and the next relay R_{k+1} as illustrated in Fig. 4. This assumption holds in the scenario of long hops where the interference with distant nodes can be neglected. Accordingly, relay R_{k+2} cannot transmit simultaneously since it will interfere with R_{k+1} that is receiving a packet from R_k. As such, R_{k+3} is the nearest relay that is allowed to transmit. Hence, for multiple-links activation, the indices k and k' of the transmitting nodes should satisfy the following relation in order to avoid any interference in the multi-hop network:

$$|k' - k| \geq 3 \quad ; \quad k, k' \in \{0, 1, \cdots, K\}. \tag{10}$$

With MLSP, either no packet or a single packet is transmitted along each of the selected links based on (8).

4) Multiple-Links Multiple-Packets (MLMP): For this setup, more than one non-interfering hop is activated with packet merging. As with MLSP, the condition in (10) should be satisfied in

order to avoid interference. Moreover, if the set of selected links comprises the k-th link, then the number of packets to be transmitted along this link is as given in (9). It is envisaged that the concurrent activation of multiple-links supporting multiple-packets each results in a more fluid flow of information between S and D which positively impacts the throughput.

While the UEs are usually equipped with a single antenna because of the size constraints especially in IoT and D2D applications, BSs might be equipped with multiple antennas. The deployment of multiple antennas will alter the specific expressions of the channel capacities in (1)-(2) and (7) without altering the DRL methodology presented in this paper. The capacity analysis of multiple-input-multiple-output (MIMO) systems is well formulated in the open literature and the extension of the presented system model to the MIMO context entails only changing the specific values of $\{C_k\}$ without affecting the associated analysis.

It is assumed that the channel is shared in a TDMA or FDMA manner implying that the transmissions from S and the relays take place in the time-frame or frequency-band, respectively, allocated to the transmitting source S. As such, the implementation of the proposed relaying strategies does entail any alterations to the medium access control (MAC) layer as compared to the conventional point-to-point communication systems. Consequently, the physical layer can be analyzed independently from the MAC layer as in [13]–[21], [26]. In this context, the relays are assumed to be independent nodes that are present in the vicinity of S and D. The relays can then assist the communications between S and D resulting in more efficient communications. This cooperation takes place in S's time-frame or frequency-band where the relays dedicate their available resources for assisting S without penalizing other users. As such, the level of fairness in the cooperative network is the same as in conventional networks even when multiple packets are merged together.

III. Problem Formulation

DRL refers to the use of deep neural networks in conjunction with RL to solve complex decision-making problems. In addition to value-based methods like DQN, there are also policy-based methods that directly output the optimal action given a state as well as the actor-critic methods that combine value-based and policy-based techniques. Being a specific algorithm within the broader category of DRL, DQN is a value-based RL algorithm that uses deep neural networks to estimate the Q-value function and improve the agent's performance. The Q-function in DQN is a type of value function that estimates the expected cumulative reward of taking an action

in a given state and following the optimal policy thereafter. The DRL based relay selection solution that we propose in this paper is based on the DQN methodology. We start by defining the state space that captures the parameters of the cooperative network as well as the action space that describes the evolution of the network. We also suggest appropriate reward functions that positively impact the throughput of the network.

The problem consists of designing a RL agent to meet the following objective: maximize the throughput of the cooperative network while minimizing the delay of packets arriving at D.

At time slot t, we denote by $\mathbf{a}_t$ the vector comprising the numbers of packets transmitted along the network's constituent links. (1): For parallel-relaying systems with no inter-relay cooperation (SPNI and MPNI), $\mathbf{a}_t$ is a $2K$-dimensional vector such that the first K components correspond to the S-R links while the remaining K components pertain to the R-D links. Denoting by $\mathbf{a}_t(k)$ the k-th component of $\mathbf{a}_t$, then $\mathbf{a}_t(k)$ corresponds to the link S-R$_k$ for $k = 1, \ldots, K$ and to the link R$_{k-K}$-D for $k = K+1, \ldots, 2K$. (2): For parallel-relaying systems with inter-relay cooperation (SPWI and MPWI), $K-1$ additional elements are appended to $\mathbf{a}_t$ where $\mathbf{a}_t(k)$ corresponds to the inter-relay link R$_{k-2K}$-R$_{k-2K+1}$ for $k = 2K+1, \ldots, 3K-1$. (3): For serial-relaying systems, $\mathbf{a}_t$ is a $(K+1)$-dimensional vector where $\mathbf{a}_t(k)$ corresponds to the number of packets transmitted along the k-th hop for $k = 1, \ldots, K+1$. We denote by K_l the number of links in the network with $K_l = 2K$, $K_l = 3K-1$ and $K_l = K+1$ for parallel-relaying with no inter-relay cooperation, parallel-relaying with inter-relay cooperation and serial-relaying, respectively.

For reliable communications, the actual number of packets transmitted along a link must not exceed the maximum number of packets that can be supported by this link. Therefore, for parallel-relaying systems, the following constraints must be satisfied:

$$\mathbf{a}_t(k) \leq \begin{cases} n_{max}(k), & k = 1, \ldots, K \text{ (S-R links);} \\ n_{max}(k-K), & k = K+1, \ldots, 2K \text{ (R-D links);} \\ n_{max}(k-2K), & k = 2K+1, \ldots, 3K-1 \text{ (R-R links)} \end{cases}, \tag{11}$$

where n_{max} is given in (3)-(4) for the S-R and R-D links and in (5)-(6) for the R-R links.

For serial-relaying systems, the constraints to be met are given by:

$$\mathbf{a}_t(k) \leq n_{max}(k) \quad ; \quad k = 1, \ldots, K+1, \tag{12}$$

where the maximum number of packets $n_{max}(k)$ that can be supported by the k-th hop is given in (8) for single-packet transmissions and in (9) for multiple-packets transmissions.

In order to maximize the throughput of the network, the number of packets delivered to D must be maximized over a sufficiently-long communication session. In parallel-relaying systems, packets are delivered to D along the links R_1-D, ..., R_K-D. Therefore, for parallel-relaying systems, the optimization problem can be formulated as follows:

$$\max_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{k=K+1}^{2K} \mathbf{a}_t(k), \tag{13}$$

which corresponds to maximizing the numbers of packets transmitted along the R-D links under the constraints in (11).

In serial-relaying systems, packets reach D through the last hop R_K-D (i.e. the $(K+1)$-th hop). Therefore, the optimization problem can be expressed as:

$$\max_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{a}_t(K+1), \tag{14}$$

under the constraints given in (12).

In addition to (11)-(12), the optimization problem must respect an additional set of constraints in order to meet the requirements of the relaying strategy. For SPNI, SPWI and SLSP, a single packet is transmitted along the single activated link. Therefore, at most one entry of vector $\mathbf{a}_t$ can be equal to 1 while all other entries will be equal to 0:

$$\sum_{k=1}^{K_l} \mathbf{a}_t(k) \leq 1. \tag{15}$$

We denote by Ψ_t the set of indices of non-zero elements of $\mathbf{a}_t$:

$$\Psi_t = \{k = 1, \ldots, K_l \mid \mathbf{a}_t(k) \neq 0\}. \tag{16}$$

For MPNI, MPWI and SLMP, only one link can be activated with the possibility of transmitting more than one packet along this link. As such, the following constraint must be met:

$$|\Psi_t| \leq 1, \tag{17}$$

implying that the cardinality of the set Ψ_t must not exceed one.

For MLSP and MLMP, the no-interference constraint in (10) must be met for all elements of Ψ_t. Moreover, for MLSP, the activated link can carry only one packet implying that the following constraint must be met as well:

$$\mathbf{a}_t(k) = 1 \quad \forall \quad k \in \Psi_t. \tag{18}$$

A summary of the optimization functions and the constraints is provided in Table I for all considered parallel-relaying and serial-relaying setups.

TABLE I

OPTIMIZATION FUNCTIONS AND CONSTRAINTS

Parallel-Relaying Scheme	Throughput	Constraints	Serial-Relaying Scheme	Throughput	Constraints
SPNI	(13)	(11), (15)	SLSP	(14)	(12), (15)
MPNI	(13)	(11), (17)	SLMP	(14)	(12), (17)
SPWI	(13)	(11), (15)	MLSP	(14)	(10), (12), (18)
MPWI	(13)	(11), (17)	MLMP	(14)	(10), (12)

IV. PROPOSED DRL-BASED BA RELAYING PROTOCOLS

A. Elements of the DRL Model

As has been previously delineated, the objective of the RL agent is to maximize the throughput of the network while minimizing the delay of packets arriving at D. The agent learns how to select the best link/links to accumulate the largest reward. The elements of RL consist of the environment, state, action, reward and agent.

1) Environment and State: The environment is the link/links selection BA cooperative network. The state is composed of two parts. (i): The buffer-state part that corresponds to the actual numbers of packets stored in the relays' buffers. (ii): The channel-state part that is related to the maximum numbers of packets that can be supported by the links in the network while respecting the reliability constraint. As such, the state can be represented by the following $(K + K_l)$-dimensional vector:

$$\mathbf{s}_t = \left[\underbrace{l_1, \ldots, l_K}_{\text{Buffer-State}}, \underbrace{n_{max}(1), \ldots, n_{max}(K_l)}_{\text{Channel-State}} \right], \tag{19}$$

where the buffer-state and channel-state components are evaluated at the corresponding time slot.

It is clear from (19) that the number of states increases exponentially with the number of relays K rendering the construction of a Q-table for Q-learning infeasible justifying the need for DRL.

2) Action: The action vector is the K_l-dimensional vector $\mathbf{a}_t$ that corresponds to the actual numbers of packets to be communicated along the constituent links. $\mathbf{a}_t$ can be equal to the all-zero vector. In this case, all the links in the network are unavailable and the system is in

outage. Following from the constraints delineated in Section III, non-zero values of $\mathbf{a}_t$ can be written under the following forms:

$$\mathbf{a}_t = \begin{cases} \mathbf{e}_k, & \text{SPNI, SPWI and SLSP;} \\ n\mathbf{e}_k, & \text{MPNI, MPWI, SLMP;} \\ \mathbf{e}_{k_1} + \mathbf{e}_{k_2} + \cdots, & \text{MLSP;} \\ n_1\mathbf{e}_{k_1} + n_2\mathbf{e}_{k_2} + \cdots, & \text{MLMP.} \end{cases}, \tag{20}$$

where $n, n_1, n_2, \ldots$ are natural integers while $\mathbf{e}_k$ stands for the k-th row of the $K_l \times K_l$ identity matrix. The integers k_i and k_j for MLSP and MLMP must satisfy the constraint in (10).

3) Reward: For parallel-relaying systems, the following reward system is adopted. (i): A negative reward (punishment) is given if the network is in outage. This discourages the agent from not triggering communications along the available links. (ii): A positive reward is given if a S-R or R-D link is activated. In order to enhance the network throughput, this reward must increase with the total number of packets transmitted in each time slot. Moreover, a higher reward should be assigned to the R-D links (as compared to the S-R links) in order to reduce the queuing delays and contribute to emptying the relays' buffers at a faster pace. We suggest the following reward function in this case:

$$r_t = \alpha \sum_{k=1}^{K_l} \mathbf{a}_t(k) + \beta(\xi/2), \tag{21}$$

where α and β are two tuning parameters. In (21), ξ stands for the hop index with $\xi = 1$ for the S-R hop and $\xi = 2$ for the R-D hop. Finally, the term $\sum_{k=1}^{K_l} \mathbf{a}_t(k)$ indicates the total number of packets exchanged in the network.

(iii): For SPWI and MPWI that support inter-relay communications, a zero reward is given if a R-R link is selected. While such selection positively contributes to balancing the buffers' loads in the subsequent time slots, activating a R-R link does not present any short-term advantage in terms of the packets' positions with respect to D in the sense that the queued packets will remain one hop away from D.

In the case of serial-relaying, we suggest a joint reward function based on the total number of packets transmitted in a time slot along with the relative positions of these packets with respect to D. In this context, the higher the number of packets and the closer they are to D, the higher the reward is. This reward encourages the transmission of more packets along the constituent links which positively impacts the throughput. Moreover, it encourages the flow of packets in the direction of D by penalizing the excessive queuing of packets at early stages in the multi-hop

network, thus, reducing the queuing delays. Finally, a negative reward (punishment) is given when the network is in outage and no links are activated. As such, for the no-outage events, the reward function can be expressed as:

$$r_t = \alpha \sum_{k=1}^{K+1} \mathbf{a}_t(k) + \beta \sum_{k=1}^{K+1} \frac{k}{K+1} \delta_{\mathbf{a}_t(k)>0}, \tag{22}$$

where α and β are two tuning parameters. In (22), the term weighed by β corresponds to the part of the reward that pertains to the relative position of the activated hop with respect to D. This part increases with the hop index k under the condition that this hop was activated and carries a non-zero number of packets (i.e. $\delta_{\mathbf{a}_t(k)>0} = 1$). Evidently, links that are not activated ($\delta_{\mathbf{a}_t(k)>0} = 0$) should not contribute to increasing the reward.

4) Handling Unfeasible Actions: An action is deemed unfeasible if the RL agent decides in favor of transmitting a number of packets along a certain link and this number exceeds the maximum allowable number of packets n_{max}. Similarly, unfeasible actions arise when the RL agent does not respect the constraints given in Table I. In this work, a-priori information will be used to compel the agent to choose feasible actions at all times. This is based on letting the agent use the previous knowledge about feasible actions so that, when training and testing, the agent selects only feasible actions. This will allow the agent to train the NN on a smaller set of actions and, hence, will have a faster convergence. Unlike the work in [27] where information about unfeasible actions as well as inconvenient actions assumed by the authors were provided to the agent, in this work, only unfeasible actions are eliminated and the agent is free to learn convenient actions to improve the performance.

5) Agent and Experiences: Q-learning is applied in this work. The agent interacts with the environment to generate a set of *experiences* where an *experience* in Q-learning is determined by the following four entities:

- *Current State* $\mathbf{s}_t$ of the environment.
- *Selected Action*: Based on the Q-values associated with the *Current State*, the ϵ-greedy strategy is applied to determine the appropriate action to be taken at this state to strike a balance between exploration and exploitation. The action $\mathbf{a}_t$ at state $\mathbf{s}_t$ is determined from:

$$\mathbf{a}_t = \begin{cases} \arg\max_{\mathbf{a}} Q_{\text{Prediction}}(\mathbf{s}_t, \mathbf{a}), & \text{with probability } 1-\epsilon; \\ \text{random selection}, & \text{with probability } \epsilon. \end{cases}, \tag{23}$$

  where ϵ is the exploration rate. The agent starts with a large value for ϵ to better explore the action space, and then ϵ is decreased to allow the agent to move to the exploitation

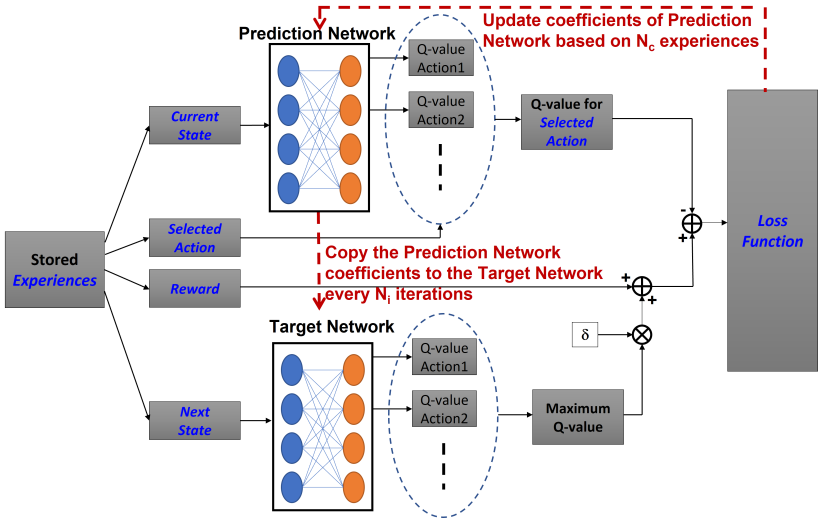

Fig. 5. DRL Block Diagram.

phase and refine the selection to the best action. In (23), the Q-values are determined from the current Q-table $Q_{\text{Prediction}}(\mathbf{s}, \mathbf{a})$ of the *Prediction Network* as will be highlighted in the next subsection.

- *Reward* r_t obtained by performing the *Selected Action* $\mathbf{a}_t$ at the *Current State* $\mathbf{s}_t$. The reward is evaluated as explained in Section IV-A3.
- *Next State*: After performing the *Selected Action* at the *Current State*, the numbers of packets stored in the relays' buffers are updated in order to reflect the changes in the network. Moreover, independent channel realisations are observed resulting in new values of the maximum numbers of packets supported by each link. This leads to the transition of the environment from the *Current State* $\mathbf{s}_t$ to the *Next State* $\mathbf{s}_{t+1}$.

B. DRL for BA Relaying

The implementation of DRL calls for two DNNs; namely, a *Prediction Network* and a *Target Network* as shown in Fig. 5. Recall that the output of a DNN is the Q-values of all actions associated with the input state. At every iteration, the *Prediction Network* has the *Current State* $\mathbf{s}_t$ as its input while the *Target Network* has the *Next State* $\mathbf{s}_{t+1}$ as its input. The DNNs are updated based on the three following steps until the system converges (the coefficients of the *Prediction Network* and *Target Network* become approximately the same).

1) Step 1 - Generate Experiences: Over N_c time slots, generate a set $\mathcal{E}$ of N_c *experiences* $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ as described in Section IV-A5. The collection of experiences (state, action, reward, next state) that the agent has observed while interacting with the environment are stored in a replay buffer. These experiences can be used to train the agent's neural network by randomly sampling batches of experiences from the replay buffer during the learning process. One of the key benefits of using replay buffers is that it helps to break the temporal correlations between experiences that can arise in RL. As such, the agent is less likely to get stuck in a local minimum or become biased towards specific experiences. Additionally, the use of a replay buffer allows the agent to learn from past experiences, even if it has moved on to exploring new parts of the state-action space. Therefore, by providing a diverse set of experiences and reducing temporal correlations, replay buffers help agents to explore the state-action space more thoroughly and learn more robust policies.

2) Step 2 - Update the Prediction Network*:* The coefficients of the *Prediction Network* are updated based on minimizing a *Loss Function* using the Adam algorithm [26]. The *Loss Function* is evaluated based on N_t randomly selected experiences from the set of experiences $\mathcal{E}$:

$$Loss = \sum_{j=1}^{N_t} \left(r_t^{(j)} + \delta \max_{\mathbf{a}} Q_{\text{Target}}(\mathbf{s}_{t+1}^{(j)}, \mathbf{a}) - Q_{\text{Prediction}}(\mathbf{s}_t^{(j)}, \mathbf{a}_t^{(j)}) \right)^2, \tag{24}$$

accounting for the discrepancy between the following quantities:

- $Q_{\text{Prediction}}(\mathbf{s}_t^{(j)}, \mathbf{a}_t^{(j)})$: the Q-value yielded by the *Prediction Network* for the *Selected Action* taken at the *Current State*.
- $r_t^{(j)} + \delta \max_{\mathbf{a}} Q_{\text{Target}}(\mathbf{s}_{t+1}^{(j)}, \mathbf{a})$: the accumulation of the immediate reward $r_t^{(j)}$ and a future discounted reward $\max_{\mathbf{a}} Q_{\text{Target}}(\mathbf{s}_{t+1}^{(j)}, \mathbf{a})$ which corresponds to the largest Q-value with respect to all actions of the *Target Network* at the *Next State*. The largest Q-value from the *Target Network* is multiplied by a discount factor δ.

3) Step 3 - Update the Target Network*:* At the end of each round encompassing N_i iterations, copy the coefficients of the *Prediction Network* to the *Target Network*.

The learning process is summarized in Algorithm 1 and a list of the DRL parameters is shown in Table II.

Note that, as with the existing BA relaying solutions in [13]–[21], [26], the proposed DRL-based protocols are centralized. In this context, the CSI and BSI must be gathered and shared with a central node that makes a decision on the links to be activated and on the numbers of packets to be communicated along these links. However, despite the obvious challenges behind

for $n = 1, \ldots, N_r$ **do**
  Reset the environment's variables
  for $i = 1, \ldots, N_i$ **do**
    Update ϵ as: $\epsilon = \max(f^{nN_i+i}, \epsilon_{min})$.
    for $j = 1, \ldots, N_c$ **do**
      For every state $\mathbf{s}_t$, get $\mathbf{a}_t$ based on (23).
      Perform $\mathbf{a}_t$ and get $\mathbf{s}_{t+1}$ and r_t.
      Store the experience $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ in $\mathcal{E}$.
    end
    Randomly choose N_t training experiences from $\mathcal{E}$.
    for $j = 1, \ldots, N_t$ **do**
      Take the j-th experience $(\mathbf{s}_t^{(j)}, \mathbf{a}_t^{(j)}, r_t^{(j)}, \mathbf{s}_{t+1}^{(j)})$.
      From *Target Network* get maximum future reward: $\max_{\mathbf{a}} Q_{\text{Target}}(\mathbf{s}_{t+1}^{(j)}, \mathbf{a})$.
      From *Prediction Network* get: $Q_{\text{Prediction}}(\mathbf{s}_t^{(j)}, \mathbf{a}_t^{(j)})$.
    end
    Get the Loss Function L_i based on (24).
    Update the *Prediction Network* based on (24).
    Clear $\mathcal{E}$.
  end
  Copy the weights of the *Prediction Network* into the *Target Network*.
end

Algorithm 1: DRL for all parallel-relaying and serial-relaying setups.

the implementation of centralized strategies, the signalling overhead in the network is judged to be limited since the elements of the state vector in (19) assume integer values thus circumventing the need for high-precision quantization of the continuous-value path gains. Naturally, the BS plays the role of the central node and, with the high computation powers available at these node, implementing the two DNNs at the BS is highly feasible from a practical point of view.

V. Numerical Results

The main performance metrics considered are the throughput and the average packet delay (APD). The throughput of a system is the average number of packets arriving at D and is measured in packets per time slot. The APD is measured by averaging the delays of all packets that arrived to D and its unit is normalized per time slot.

Simulations are carried out over Rayleigh block-fading channels where the channel coefficients are generated independently from one time slot to another. For the DRL implementation, we fix

TABLE II

SUMMARY OF DRL PARAMETERS

N_r	Number of rounds for RL
N_i	Number of iterations per round
N_c	Number of collected experiences per iteration
N_t	Training batch size
δ	Discount factor
ϵ	Exploration rate
f	Decay factor
$\mathcal{E}$	Set of saved experiences per iteration

the discount factor to 0.8 and the learning rate of the Adam algorithm to 0.01. For implementing the *Prediction Network* and *Target Network*, 3-layers DNNs are deployed. The first two layers comprise 64 neurons each with leaky-Relu activation while the number of neurons in the last layer is equal to the number of actions with Relu activation. Finally, we fix $N_r = 25$, $N_i = 50$, $N_c = 200$, $N_t = 32$, $f = 0.999$ and $\epsilon_{min} = 0.1$. Fixing the DRL parameters as in Table II, the performance of the cooperative network can be fully determined from the target rate r_0, the buffer size L and the average channel gains $\Omega_1, \ldots, \Omega_{K_l}$ of the network's links. In what follows, we fix $r_0 = 1$ bit per channel use while the parameters L and $\Omega_1, \ldots, \Omega_{K_l}$ will be varied in the different simulation setups. Assuming a path loss exponent of 2 and a loss of 30 dB at a reference distance of 1 km, the average channel gains can be related to the link distances by $10\log_{10}(\Omega_k) = 30 - 20\log_{10}(d_k)$ where d_k stands for the length of the k-th link.

This section includes comparisons of the proposed schemes with the benchmark parallel-relaying and serial-relaying schemes in [19], [26] and [21], respectively. The scheme in [26] disregards the packets whose delays exceed a tolerated ceiling while the delays of all packets arriving at D are included in the APD evaluations in our work. As such, the DRL-based parallel-relaying scheme in [26] becomes equivalent to the proposed SPNI scheme if the delay constraint is relaxed. Therefore, the throughput and APD curves presented for the SPNI scheme also capture the performance levels realized by the benchmark scheme in [26]. On the other hand, the threshold-based parallel-relaying scheme in [19] constitutes the state-of-the-art BA relaying protocol that is capable of achieving a broad range of tradeoffs between reliability (or equivalently throughput) and delay. These tradeoffs can be achieved by fixing threshold levels of either 0 or

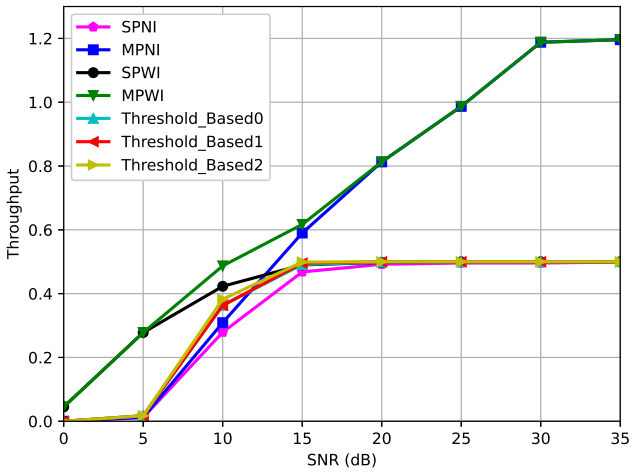

Fig. 6. Throughput of a 2-relay BA parallel-relaying network.

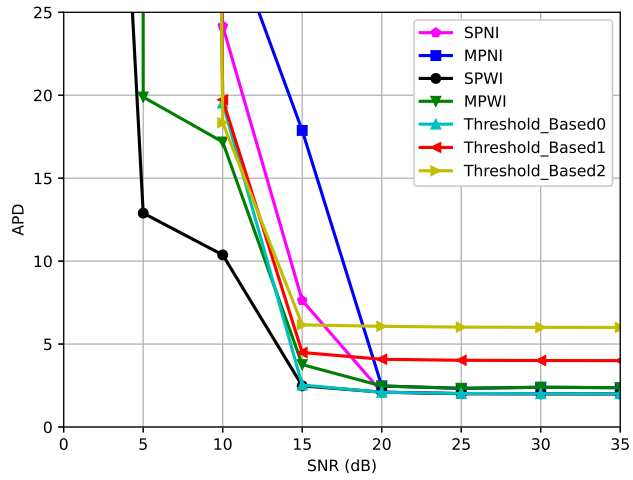

Fig. 7. APD of a 2-relay BA parallel-relaying network.

1 at the K relays where each one of these levels determines the role of the corresponding relay in the cooperation environment. In the presented numerical results, the scheme in [19] will be labeled as "Threshold_Based$N$" where N stands for the number of threshold levels that are equal to 0 allowing to achieve a diversity order of $K+N$ with an asymptotic APD value of $2N+2$ [19]. For the serial-relaying setup, the scheme in [21] constitutes the reference BA scheme. This scheme assigns weights to the $K+1$ hops and activates the hop with the largest weight. The

weight of the first hop is denoted by the parameter s that controls the tradeoff between reliability and delay. In particular, the achievable asymptotic APD is equal to $2K + (s-1)K(K+1)$ and the diversity order is 1, K and $K+1$ for $s = 1$, $s = L$ and $1 < s < L$, respectively. Finally, to the authors' best knowledge, there are no DRL-based serial-relaying solutions in the literature.

Fig. 6 and Fig. 7 show the variations of the throughput and APD, respectively, as a function of the SNR for a 2-relay BA parallel-relaying network with $L = 8$, $(\Omega_1, \Omega_2) = (1, 0.2)$ and $(\Omega_1', \Omega_2') = (0.2, 1)$. Results show that restricting the transmissions to a single packet at a time severely undermines the throughput and more than a two-fold improvement in the throughput at high SNR can be realized by transmitting more than one packet along high-quality links. These results further support the validity of the relaying solutions presented in this work. The improvements that follow from activating the R-R links manifest at low SNR since, at high SNR, the probability that none of the S-R and R-D links is available is very small. At a SNR of 10 dB, SPWI and MPWI activate the inter-relay links around 17% and 22% of the time, respectively, resulting in visible improvements in the throughput following from Fig. 6. Results in Fig. 7 show that the aforementioned throughput improvements are associated with significant reductions in the APD at low-to-average SNRs. In fact, allowing for the flow of packets from relay R_1 to relay R_2, that has a better channel quality with D, circumvents the excessive queuing of the packets at R_1 which positively contributes to reducing the delays. Results in Fig. 6 show that none of the benchmark schemes SPNI, Threshold_Based0, Threshold_Based1 and Threshold_Based2 in [19], [26] can increase the throughput of the two-relay network beyond the value of 0.5 even for large SNR values. This highlights on the advantages of the proposed MPNI and MPWI schemes that allow for the transmission of multiple packets along the network's links allowing to achieve asymptotic throughput values in the order of 1.2 packets per time slot. Fig. 7 shows that these throughput improvements of the MPNI and MPWI schemes are associated with advantageously small APD values that tend to the best previously reported asymptotic APD of 2 achieved by the delay-prioritizing scheme Threshold_based0 in [19].

Fig. 8 and Fig. 9 show the variations of the throughput and APD, respectively, as a function of the SNR for a 6-relay BA parallel-relaying network with $L = 8$, $(\Omega_1, \ldots, \Omega_6) = (4, 3.5, 3, 2.5, 2, 1.5)$ and $(\Omega_1', \ldots, \Omega_6') = (1, 1.5, 2, 2.5, 3, 3.5)$. Results in Fig. 8 and Fig. 9 show that the activation of the inter-relay links does not present any performance advantage in the considered simulation scenario where the network comprises a large number of relays ($K = 6$). In fact, the probability that all $2K = 12$ S-R and R-D links suffer from outage is very small

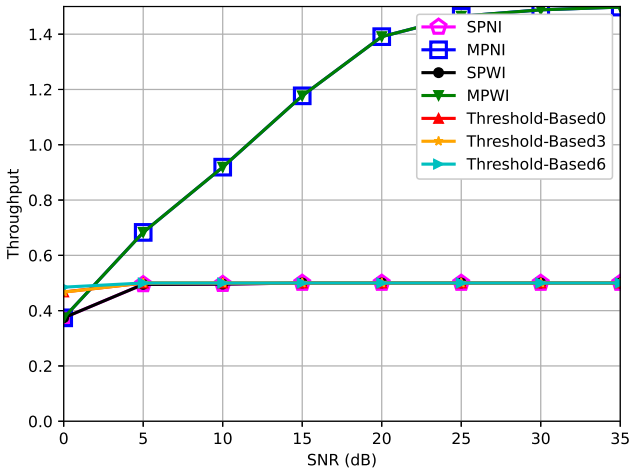

Fig. 8. Throughput of a 6-relay BA parallel-relaying network.

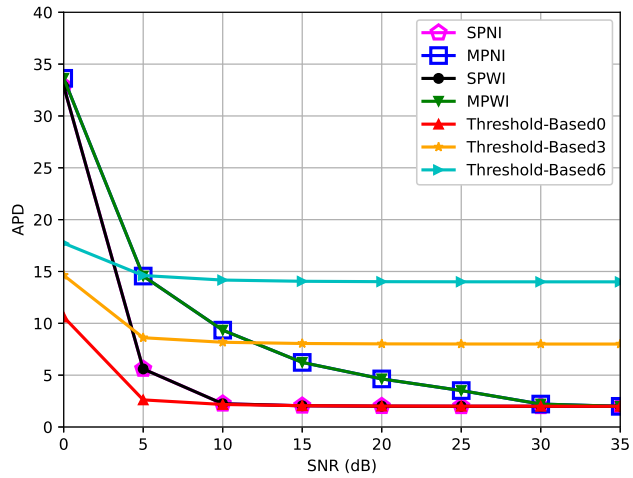

Fig. 9. APD of a 6-relay BA parallel-relaying network.

even at small SNR values. For example, even if the outage probability along each link is as large as 10^{-1}, then the probability that all links in the network are in outage will scale as 10^{-12} where this probability is very small and, hence, can be neglected. Therefore, from figures 6-9, we can conclude that inter-relay cooperation is the most useful at low-to-average values of the SNR for networks that do not comprise a large number of relays. The APD variations for the 2-relay and 6-relay networks in Fig. 7 and Fig. 9, respectively, demonstrate the efficiency of the

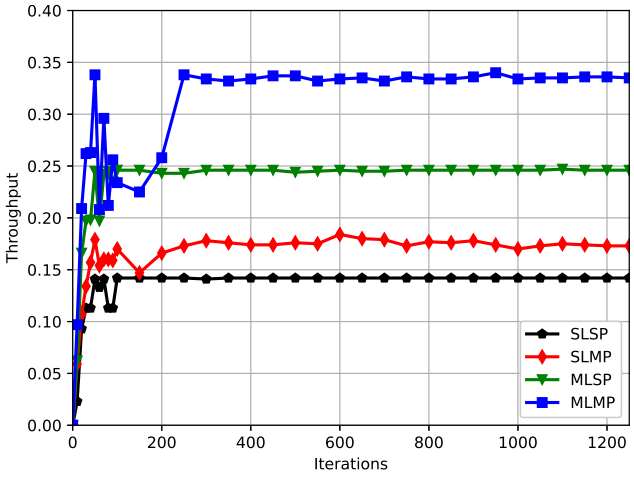

Fig. 10. Throughput of a 6-relay BA serial-relaying network.

RL agent since, at high SNRs, the APDs converge to the minimum possible value of 2 since a packet cannot be delivered from S to D in less than two time slots (one slot for the S-R hop and a second slot for the R-D hop). Results in Fig. 8 show that the MPNI and MPWI schemes are capable of achieving a very large throughput that is three times higher than that of SPNI and SPWI that restrain the transmission to a single packet along each link. However, from Fig. 9, the transmission of multiple packets results in an increase in the APD at low-to-average SNR values. In fact, consider the case where the quality of the link S-R_k is good and the quality of the link R_k-D is poor (as is the case for relay R_1 in the considered simulation setup where $\Omega_1 = 4$ and $\Omega_1' = 1$). In this case, a large number of packets can be supported by the link S-R_k. However, the packets that reach R_k will be eventually queued for longer times since the link R_k-D cannot support the transmission of a large number of packets following from its poor channel conditions. This queuing results in increased delays at low-to-average SNRs. On the other hand, at large SNRs, all channel conditions are favorable and the delays of all 4 setups converge to the optimal value of 2. Conclusions pertaining to the comparison with the benchmark schemes in [19], [26] are analogous to those observed in Fig. 6 and Fig. 7. In particular, MPNI and MPWI concurrently increase the throughput and reduce the APD where the performance improvements are more significant at large SNR values. The asymptotic APD value of 2 achieved by SPNI, SPWI, MPNI and MPWI is much smaller than the values of 8 and 14 achieved by

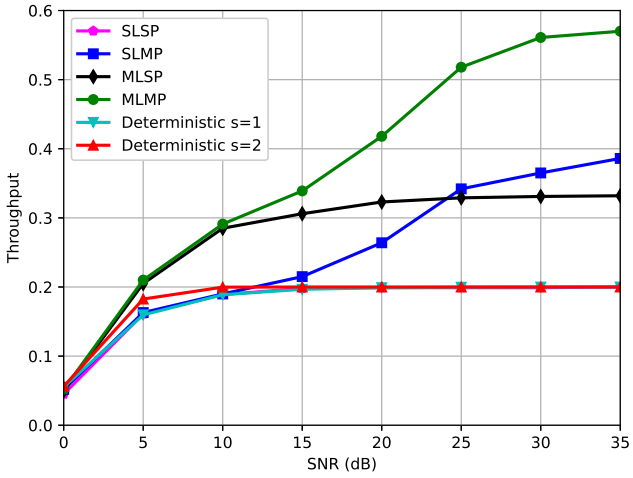

Fig. 11. Throughput of a 4-relay BA serial-relaying network.

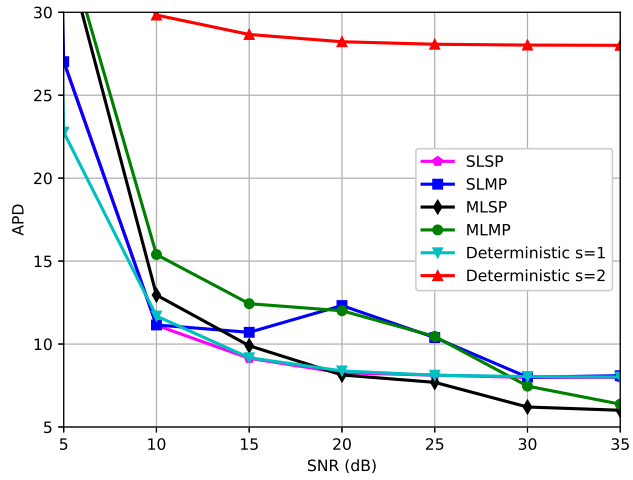

Fig. 12. APD of a 4-relay BA serial-relaying network.

Threshold_Based3 and Threshold_Based6.

In order to analyze the convergence of the DRL algorithm, Fig. 10 shows the variation of the throughput as a function of the number of iterations for a 6-relay serial-relaying system at a SNR of 30 dB. We fix $L = 10$ and we consider symmetrical hops with $\Omega_k = 12$ for $k = 1, \ldots, 7$. Results show that the design of the DRL framework manifests fast convergence for all setups. MLMP takes a longer time to converge since the number of actions is the largest. The benchmark

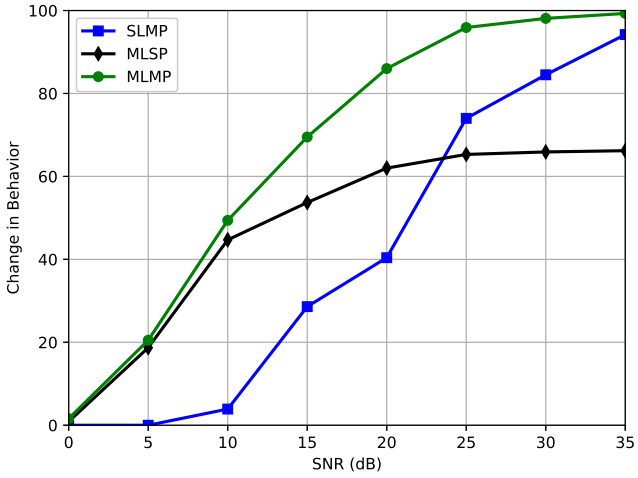

Fig. 13. Change in behavior of a 4-relay BA serial-relaying network.

SLSP scheme has the lowest throughput and this throughput can be significantly improved by allowing for multiple-packet transmissions and/or multiple-link activation. In this context, the multiple-link activation has a more predominant effect on the performance where the MLSP and MLMP schemes achieve the highest values of the throughput. In fact, SLSP and SLMP activate only one hop out of the seven hops which severely degrades the throughput of the end-to-end communication. The performance improvements are directly related to advantageous deviations from the standard SLSP operation. For SLMP, links carry more than one packet 30% of the time. For MLSP, more than one link is activated 75% of the time. Finally, for MLMP, multiple links are activated or multiple packets are transmitted around 88% of the time.

Fig. 11, Fig. 12 and Fig. 13 show the performance of a serial-relaying network with $K = 4$, $L = 10$ and $\Omega_1 = \cdots = \Omega_5 = 12$. The performance of the benchmark deterministic serial-relaying BA scheme in [21] is also shown and labeled as "Deterministic" where different values of the performance-controlling parameter s are considered. From Fig. 11, it can be observed that allowing for the activation of multiple links results in throughput improvements for all values of the SNR. In this context, MLSP outperforms SLSP, and MLMP outperforms SLMP at all SNRs. On the other hand, allowing for the transmission of multiple packets results in throughput improvements only for average-to-large values of the SNR. For example, at SNRs below 10 dB, SLMP results in the same throughput as SLSP; similarly, MLMP and MLSP achieve the same

throughput. In fact, at such low SNRs, the channel conditions along the constituent hops are poor implying that these hops cannot support the reliable transmission of more than packet and n_{max} cannot exceed one. Results in Fig. 11 show that the throughputs of the single-packet schemes (SLSP and MLSP) converge to a limit. On the other hand, the throughput keeps increasing with the SNR for the multiple-packets schemes (SLMP and MLMP). In fact, the throughput of SLMP and MLMP will saturate if all hops manifest a good enough quality to support the maximum possible number of packets of $L = 10$. Evidently, this can take place at impractical excessively large values of the SNR. Therefore, it can be concluded that allowing for the transmission of multiple packets will always lead to throughput improvements for practical SNR values. Results in Fig. 12 show that the multiple-links schemes MLSP and MLMP manifest the smallest high-SNR delays since, for the considered 4-relay system, nodes S and R_3 as well as nodes R_1 and R_4 can transmit simultaneously. Fig. 13 shows the variations of the percentage of change in behavior with respect to the SNR. The change in behavior is defined as the ratio of actions taken that allow for multiple packets and/or multiple links for a given setup. In other words, the curves in Fig. 13 capture the improvements compared to the standard SLSP setup. For the SLMP scheme, no change in behavior is observed below 5 dB since the links with small channel capacities cannot support multiple packets. However, the improvements increase rapidly with the SNR where, at a SNR of 30 dB, the activated links carry more than one packet around 85% of the time. For the multiple-links schemes MLSP and MLMP, the change in behavior is visible for all SNRs where these schemes benefit from the capability of simultaneously communicating over more than one link. For example, for the MLSP scheme at a SNR as low as 5 dB, the RL agent is activating more than one link around 20% of the time. At high SNRs, MLMP benefits from the high quality of the links to simultaneously transmit multiple packets over multiple links which results in the highest deviation from the standard SLSP operation. For example, at a SNR of 35 dB, the MLMP RL agent is almost never selecting a single link to carry a single packet. Results in Fig. 11 show that the throughput of the scheme in [21] with a 5-hop network cannot exceed 0.2 while the proposed SLMP, MLSP and MLMP schemes can achieve much bigger throughput values. On the other hand, Fig. 12 shows that the scheme in [21] with $s = 2$ results in drastically large values of the APD while the delay-prioritizing variant with $s = 1$ results in delay values that are comparable to those realized by the proposed schemes implying that the achievable throughput gains do not penalize the delay performance of the multi-hop network.

VI. Conclusion and Future Work

Conventional BA relaying networks are designed around two basic assumptions: a single link is activated and a single packet is transmitted along this link. This article proposed to improve the performance of the existing BA systems by relaxing these highly restrictive assumptions through the application of DRL methods to solve the complicated throughput maximization problem. Simulations demonstrated significant performance gains for the parallel and serial relaying setups. The adequate design of the reward function ensured fast convergence speeds despite the very large number of states and actions involved in the learning process.

The proposed DRL-based BA relaying schemes are centralized in the sense that a central node, for example S, must collect all buffer state and channel state information to make a decision on the link/links to be activated. Future work must consider the distributed implementation of such relaying protocols. This decentralisation is particularly pertinent to serial-relaying networks in order to reduce the signalling overhead since any central node cannot be directly reached by the remaining nodes in one hop. Finally, the proposed schemes can be generalized to more complicated networks with energy-harvesting, multiple users and/or multiple antennas.

References

[1] C. Hoymann, W. Chen, J. Montojo, A. Golitschek, C. Koutsimanis, and X. Shen, "Relaying operation in 3GPP LTE: challenges and solutions," *IEEE Communications Magazine*, vol. 50, no. 2, pp. 156–162, 2012.

[2] P.-H. Chou, "Unlicensed Band Allocation for Heterogeneous Networks," *IEICE Trans. on Commun.*, vol. 103, no. 2, pp. 103–117, Feb. 2020.

[3] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1617–1655, Third Quarter 2016.

[4] C. V. Anamuro, N. Varsier, J. Schwoerer, and X. Lagrange, "Distance-aware relay selection in an energy-efficient discovery protocol for 5G D2D communication," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4379–4391, July 2021.

[5] M. A. G. Ghasri and A. M. A. Hemmatyar, "A new dynamic optimal M2M RF interface setting in relay selection algorithm (DORSA) for IoT applications," *IEEE Access*, pp. 5327–5342, Jan. 2022.

[6] B. Li, S. Zhao, R. Miao, and R. Zhang, "A survey on unmanned aerial vehicle relaying networks," *IET Communications*, vol. 15, no. 10, pp. 1262–1272, Aug. 2021.

[7] B. Ji, Y. Han, Y. Wang, D. Cao, F. Tao, Z. Fu, P. Li, and H. Wen, "Relay cooperative transmission algorithms for IoV under aggregated interference," *IEEE Trans. on Intelligent Transportation Systems*, 2021, accepted for publication.

[8] A. Al Amin and S. Y. Shin, "Capacity analysis of cooperative NOMA-OAM-MIMO based full-duplex relaying for 6G," vol. 10, no. 7, pp. 1395–1399, July 2021.

[9] T.-Y. Wang, P.-H. Chou and W.-J. Huang, "Relay misbehavior detection for robust diversity combining in cooperative communications," *IEEE Signal Processing and Signal Processing Education Workshop*, pp. 184–189, 2015.

[10] P.-H. Chou, "Modeling the unlicensed band allocation for LAA with buffering mechanism," *IEEE Commun. Letters*, vol. 23, no. 3, pp. 526–529, March 2019.

[11] N. Zlatanov, A. Ikhlef, T. Islam, and R. Schober, "Buffer-aided cooperative communications: Opportunities and challenges," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 146–153, April 2014.

[12] N. Nomikos, T. Charalambous, I. Krikidis, D. N. Skoutas, D. Vouyioukas, M. Johansson, and C. Skianis, "A survey on buffer-aided relay selection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1073–1097, Second Quarter 2016.

[13] I. Krikidis, T. Charalambous, and J. S. Thompson, "Buffer-aided relay selection for cooperative diversity systems without delay constraints," *IEEE Trans. Wireless Commun.*, vol. 11, no. 5, pp. 1957–1967, May 2012.

[14] M. Oiwa, C. Tosa, and S. Sugiura, "Theoretical analysis of hybrid buffer-aided cooperative protocol based on max–max and max–link relay selections," *IEEE Trans. Veh. Technol.*, vol. 65, no. 11, pp. 9236–9246, Nov. 2016.

[15] A. A. M. Siddig and M. F. M. Salleh, "Balancing buffer-aided relay selection for cooperative relaying systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 8276–8290, Sep. 2017.

[16] B. Manoj, R. K. Mallik, and M. R. Bhatnagar, "Performance analysis of buffer-aided priority-based max-link relay selection in DF cooperative networks," *IEEE Trans. Commun.*, vol. 66, no. 7, pp. 2826–2839, July 2018.

[17] S. Luo and K. C. Teh, "Buffer state based relay selection for buffer-aided cooperative relaying systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5430–5439, Oct. 2015.

[18] P. Xu, Z. Ding, I. Krikidis, and X. Dai, "Achieving optimal diversity gain in buffer-aided relay networks with small buffer size," *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 8788–8794, Oct. 2015.

[19] S. El-Zahr and C. Abou-Rjeily, "Threshold based relay selection for buffer-aided cooperative relaying systems," *IEEE Trans. Wireless Commun.*, vol. 2, no. 9, pp. 6210–6223, Sep. 2021.

[20] B. Manoj, R. K. Mallik, and M. R. Bhatnagar, "Buffer-aided multi-hop DF cooperative networks: A state-clustering based approach," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 4997–5010, 2016.

[21] S. El-Zahr and C. Abou-Rjeily, "Buffer state based relay selection for half-duplex buffer-aided serial relaying systems," *IEEE Trans. Commun.*, vol. 70, no. 6, pp. 3668–3681, June 2022.

[22] S. Dong, P. Wang, and K. Abbas, "A survey on deep learning and its applications," *Elsevier Computer Science Review*, vol. 40, p. 100379, May 2021.

[23] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, Fourth Quater 2019.

[24] W. Chen, X. Qiu, T. Cai, H.-N. Dai, Z. Zheng, and Y. Zhang, "Deep reinforcement learning for internet of things: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1659–1692, Third Quarter 2021.

[25] S. Dong, Y. Xia, and T. Peng, "Network abnormal traffic detection model based on semi-supervised deep reinforcement learning," *IEEE Transactions on Network and Service Management*, vol. 18, no. 4, pp. 4197–4212, Oct. 2021.

[26] C. Huang, G. Chen, and Y. Gong, "Delay-constrained buffer-aided relay selection in the internet of things with decision-assisted reinforcement learning," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 10 198–10 208, June 2021.

[27] C. Huang, G. Chen, Y. Gong, P. Xu, Z. Han, and J. A. Chambers, "Buffer-aided relay selection for cooperative hybrid NOMA/OMA networks with asynchronous deep reinforcement learning," *IEEE J. Select. Areas Commun.*, vol. 39, no. 8, pp. 2514–2525, Aug. 2021.

[28] C. Huang, G. Chen, Y. Gong, and Z. Han, "Joint buffer-aided hybrid-duplex relay selection and power allocation for secure cognitive networks with double deep Q-network," *IEEE Trans. Cognitive Commun. and Networ.*, vol. 7, no. 3, pp. 834–844, Sep. 2021.