Leveraging Reinforcement Learning for User Pairing in Full Duplex Networks

João Rafael Barbosa de Araújo e Francisco Rafael Marques Lima

Abstract— **In this article we employ a reinforcement learning solution called Upper Confidence Bound (UCB) over the framework of Multi-Armed Bandit (MAB) to solve User Equipment (UE) pairing problem in Full Duplex (FD) network. In the context of the total data rate maximization problem, our proposed solution is capable of learning the best UE pair iteratively by exploring and exploiting the solution space. By the presented simulation results, we show that our proposed algorithm is more robust to the absence of knowledge about inter-UE Channel State Information (CSI). In the complete absence of CSI about inter-UE channel gains, our proposed solution overperforms the Maximum Rate (MR) solution by 26%.**

Keywords— **Reinforcement Learning, Full-Duplex, Multi-armed bandit.**

I. Introduction

Mobile communication sector is in constant evolution and it is one of the most important industries nowadays since it is directly related to the development of several other businesses. However, the demands over mobile networks increase at a rapid pace. According to Ericsson Mobility Report [1], there will be 8.9 billion of mobile subscriptions in 2025 (without accounting Internet of Things (IoT) devices). When mobile traffic is regarded, from the first quarter of 2019 to the first quarter of 2020, mobile data traffic grew 56%.

In order to cope with the increased number of subscriptions and mobile data traffic, more frequency bands should be available in order to improve system capacity. Unfortunately, the release of new frequency bands or refarming of existing ones is a slow and intricate process in nationwide and worldwide scopes. In this sense, new techniques that are able to improve spectral efficiency have been studied in the literature [2]. In this article, we deal with in-band Full Duplex (FD) technology.

FD is expected to be the next step towards efficient use of spectrum resources. With FD, transceivers are able to transmit and receive signals in the same time-frequency resource, thus boosting spectral efficiency by providing a twofold performance gain compared to typical Half Duplex (HD) networks [2]. The interest in FD technology has arisen last years thanks to the advances in the capacity to attenuate the so called Self Interference (SI), i.e., the signal transmitted by one device that appears as interference to its own receiver.

Interestingly, FD gains can be attained even in point-to-multipoint scenarios with FD-capable Base Station (BS) and HD User Equipments (UEs). In this case, the same time-frequency resource can be used in downlink (from BS to UE) and in uplink (from UE to BS). This is a very appealing solution in an initial deployment of FD technology since the HD mobile terminals can remain with cheap hardware and processing capabilities whereas most of the complexity is left to the FD-capable BS.

However, in this setup, some problems should be mitigated. Besides SI that must be canceled at the receiver of the BS, the uplink UE generates interference to the other UE in downlink direction; the so called Co-Channel Interference (CCI). In order to combat the CCI, the proper choice of uplink and downlink mobile terminals to share the same resource is of utmost importance. In this article we define the choice of uplink and downlink UEs to transmit on FD mode as UE pairing.

Conventionally, in order to mitigate CCI, the BS need to know specific information about UE's geographical positions or Channel State Information (CSI) of inter-UE links. In order to make this available, control messages should be exchanged between UEs and BS, which consumes precious time-frequency resources that otherwise could be used to transmit data traffic.

Machine Learning (ML) have been successfully applied in many areas such as computer vision and autonomous systems. Recently, ML algorithms have been applied to telecommunications problems including network management and physical layer optimizations [3]. In this article, we intend to show the potential of the use of ML algorithms to the UE pairing in FD point-to-multipoint connections in order to solve the total data rate maximization problem.

The remaining of this article is organized as follows. In Section II we present the state of the art related to this research topic. In Section III we present the assumed system model as well as the mathematical formulation of the studied problem. Sections IV and V present the proposed solution and numerical results, respectively. Finally, the main conclusions of our article are depicted in section VI.

II. Literature review

Radio resource allocation for FD networks has been studied in the literature with different approaches and problem objectives. The interested reader can see [2] for a deeper

João Rafael B. de Araújo and F. Rafael M. Lima are with the Post-Graduation Program on Computer and Electrical Engineering, Federal University of Ceará - Sobral Campus, Sobral-Brazil, e-mail: rafael.acurcio@gmail.com. F. Rafael M. Lima is also with the Wireless Telecom Research Group (GTEL), Federal University of Ceará - Pici campus, Fortaleza-Brazil, e-mail: rafaelm@gtel.ufc.br. The research of João R. B. Araújo was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior CAPES - Brasil.

survey on different problems in FD networks. In this article, our focus is on the total data rate maximization problem in point-to-multipoint scenario.

Some articles have studied the joint resource allocation problem of UE pairing, subcarrier assignment and power allocation for total data rate maximization [4, 5, 6]. In [4], those problems were decomposed into three subproblems to reduce complexity while in [5] the authors provide the necessary conditions for optimality and proposed a suboptimal solution. To tackle this complex problem more efficiently, the authors in [5] formulate the joint problem as a three-sided matching problem. In [6], the same approach to decompose the joint problem into subproblems is employed with an iterative algorithm. The main difference is that, differently of [4, 5], the authors in [6] consider the unrealistic assumption of FD UEs, which imposes a high complexity burden on the UE side.

In [7], the authors solved the joint problem in a different scenario setup: heterogeneous networks. In this case, the authors also considered the UE-BS association problem. In the new problem, firstly, the integer variables are relaxed and then, the resulting problem is solved with interior-point method. In all aforementioned works, the authors employed classical optimization techniques to obtain the solutions. Furthermore, they assumed perfect CSI in all links in an FD network; including inter-UE links. According to the related literature, the inter-UE CSI acquisition is performed by means of pilot signals that are sent by all UEs in specific orthogonal time-frequency channels. After estimating the inter-UE CSI, those measurements should be sent to the BS by means of control channels so as to the BS take centralized decisions about subcarrier assignment, power allocation and, mainly, UE pairing. Therefore, full CSI knowledge in FD network is an assumption difficult to hold in practice and may drain all the theoretical performance gains of FD.

As previously mentioned, in this article we leverage ML solutions to solve UE pairing problem in FD networks. Particularly, we are interested in this article in a specific category of ML called Reinforcement Learning [8]. Reinforcement Learning is a process where an algorithm interacts with an environment through an agent and evaluate its actions by a reward function. Based on the rewards, a strategy can be derived to find/improve a policy.

Our focus in this article is in the application of Multi-Armed Bandit (MAB) problem; an ML framework that belongs to the class of reinforcement learning. MAB is a reinforcement learning problem where a fixed limited set of actions must be performed in a way that maximizes the expected gain [8]. For each action, a value reward is given and the distribution of the value rewards is not initially known. Trying to maximize expected gain by using current system estimates is known as a greedy strategy, where the current estimated knowledge of the system is *exploited*. Another option is to *explore* by choosing non-optimal actions to better estimate the value rewards for some chosen actions. Therefore, there is a trade-off between *exploration* vs *exploitation*.

As far as we know, few articles have applied MAB in the context of FD networks. In [9] the authors consider a scenario where Unmmaned Aerial Vehicle (UAV) is used as a relay

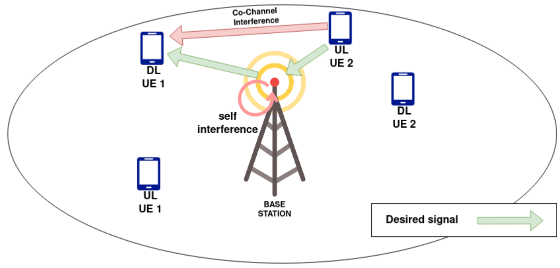


Fig. 1: Illustration of an FD-capable BS and UEs in a point-to-multipoint scenario.

between a BS and terrestrial vehicles. The UAVs are FD-capable but MAB is employed to solve the problem of UAV positioning so as to maximize the total data rate. In [10] the authors consider FD stations but in the context of Carrier Sense Multiple Access (CSMA) networks. MAB is applied to maximize total data rate by adjusting transmit power and carrier sense threshold.

In summary, the main contribution of our article is the use of MAB to perform UE pairing. Differently of the previous works, our proposal does no rely on the perfect knowledge of inter-UE CSI. Indeed, our proposed algorithm is able to interact with the environment and, after a convergence period, learn an efficient policy to perform UE pairing.

III. System Model

We assume a single-cell system with an FD-capable BS and multiple HD UEs in a point-to-multipoint wireless network. In a given Transmission Time Interval (TTI), the BS is capable of transmitting information to a UE, hereby defined as Downlink (DL) UE, and receiving information from another UE, hereby defined as Uplink (UL) UE. As previously explained, the uplink is impaired by SI whereas the downlink direction is impaired by the CCI generated by the UL UE. Figure 1 illustrates this scenario.

We assume that $\mathcal{I} = \{1, \cdots, I\}$ is the set of DL UEs while $\mathcal{J} = \{1, \cdots, J\}$ is the set of UL UEs. Without loss of generality, we assume in this article that the number of DL and UL UEs is the same, i.e., $I = J$. Moreover, we consider an Orthogonal Frequency Division Multiple Access (OFDMA) network where a Resource Block (RB) is defined as a time-frequency grid composed of a number of Orthogonal Frequency Division Multiplexing (OFDM) symbols in the time domain and a given number of subcarriers in the frequency domain. We define $\mathcal{F} = \{1, \cdots, F\}$ as the set of available RBs.

The adopted channel model takes into account the distance-dependent path loss, shadowing and fast fading. Before presenting the involved variables, we highlight that all variables are time-dependent, i.e., they depend on the current TTI, t. For the sake of notational simplicity, we omit the time dimension and will only include it when it is relevant to the context. We define $g^{\mathrm{d}}_{i,f}$ as the channel gain between the BS and DL UE i on RB f in downlink. Furthermore, $g^{\mathrm{u}}_{j,f}$ is defined as the channel gain between UL UE j and the BS on RB f in uplink. The channel gain between UL UE j and DL UE i on RB f is represented by $g_{i,j,f}$.

We assume that a constant transmit power per RB is applied in both downlink an uplink. So, we define p^{d} and p^{u} as the transmit power per RB on the BS and on UEs, respectively. As previously explained, the BS is capable of canceling part of the SI. In our model, we assume that the remaining SI after cancellation is given by $\beta \cdot p^{\mathrm{d}}$ for a given RB where β accounts for the capacity of mitigating SI.

We consider a dynamic scenario where at each TTI, the BS performs UE pairing, i.e., chooses a pair of UL and DL UEs to transmit and receive information, respectively. We consider an OFDMA/Time Division Multiple Access (TDMA) network where the chosen pair gets assigned the whole bandwidth in a given TTI. According to those definitions and assuming that the pair DL UE i and UL UE j was chosen in a specific TTI, the experienced Signal to Interference plus Noise Ratio (SINR) in downlink and uplink in RB f are given by

$$\gamma_{i,j,f}^{\mathrm{d}} = \frac{p^{\mathrm{d}} \cdot g_{i,f}^{\mathrm{d}}}{\sigma^2 + p^{\mathrm{u}} \cdot g_{i,j,f}}, \text{ and } \gamma_{j,f}^{\mathrm{u}} = \frac{p^{\mathrm{u}} \cdot g_{j,f}^{\mathrm{u}}}{\sigma^2 + p^{\mathrm{d}} \cdot \beta}, \quad (1)$$

respectively, where σ^2 is the noise power in the bandwidth of an RB.

We assume a link adaptation functionality where the transmit data rate is adapted according to the channel quality by selecting different combinations of modulation order and channel coding rate, i.e., Modulation and Coding Scheme (MCS). Consider a function $e\left(m, \gamma\right)$ that has as input an MCS index m that belongs to the set $\mathcal{M} = \{1, \cdots, M\}$ and an SINR γ, and returns the block error probability. M is the number of available MCSs. Furthermore, we assume that the raw data rate transmitted when using MCS m is w_m. According to this, the effective data rate when transmitting with MCS m in an RB for a received SINR of γ is given by $r^{\mathrm{e}}\left(m, \gamma\right) = w_m \cdot \left(1 - e\left(m, \gamma\right)\right)$.

The assumed link adaptation in this work is based on throughput maximization. Therefore, for a given estimated SINR, γ, on each RB, the selected MCS is the one that maximizes the effective data rate, i.e.,

$$m^* = \arg \max_{m \in \mathcal{M}} \{r^{\mathrm{e}}\left(m, \gamma\right)\}. \quad (2)$$

Based on the previous definitions, assume henceforth that function $g\left(\gamma\right)$ returns the effective data rate based on the just described link adaptation method. Note that γ is given by equation (1). Assuming that a given pair DL UE i and UL UE j was selected to transmit on TTI t, then, the transmitted effective data rate on downlink and uplink[1] are, respectively,

$$r_{i,j}^{\mathrm{tot,d}}\left(t\right) = \sum_{f \in \mathcal{F}} g\left(\gamma_{i,j,f}^{\mathrm{d}}\left(t\right)\right) \text{ and } r_{i,j}^{\mathrm{tot,u}}\left(t\right) = \sum_{f \in \mathcal{F}} g\left(\gamma_{j,f}^{\mathrm{u}}\left(t\right)\right). \quad (3)$$

As motivated in section II, we also evaluate the impact of imperfect CSI on inter-UE links. Therefore, when perfect CSI is not known at the BS, the link adaptation illustrated in equation (2), as well as other functionalities at the BS, are performed with an estimate of $g_{i,j,n}$, i.e., $\hat{g}_{i,j,n}$. In this work, we model the CSI knowledge of $g_{i,j,n}$ at the BS, $\hat{g}_{i,j,n}$, in different levels: (CSI1) or perfect CSI where $\hat{g}_{i,j,n} = g_{i,j,n}$; (CSI2) where $\hat{g}_{i,j,n}$ is composed by path loss and shadowing components of $g_{i,j,n}$ whereas fast fading is left out; (CSI3) where $\hat{g}_{i,j,n}$ is composed only by path loss component of $g_{i,j,n}$ whereas shadowing and fast fading are not known; finally, (CSI4) where $\hat{g}_{i,j,n} = 0$, i.e., the case when the interference between UEs is completely ignored.

[1]Note that in these equations the time dimension is explicit for convenience.

Our objective in this work is to maximize the long-term data rate over a time horizon of T TTIs by performing UE pairing at each TTI. Assume that $x_{i,j,t}$ is equal to 1 if the pair DL UE i and UL UE j is selected by the BS at TTI t, and 0 otherwise. This can be mathematically formulated as

$$\max_{x_{i,j,t}} \left\{ \sum_{t=1}^{T} x_{i,j,t} \cdot \left(r_{i,j}^{\mathrm{tot,d}}\left(t\right) + r_{i,j}^{\mathrm{tot,u}}\left(t\right) \right) \right\}, \quad (4a)$$

$$\text{s.t.} \sum_{\forall i \in \mathcal{I}} \sum_{\forall j \in \mathcal{J}} x_{i,j,t} \leq 1, \;\; \forall t \in \{1, \cdots, T\}. \quad (4b)$$

IV. Proposed Solution

In order to solve problem (4) we propose the use of MAB framework. Basically, in MAB, at each iteration an *agent* has to choose one specific *action* from a list of available ones (action set) in order to maximize an expected (long-term) gain. After taking a decision, the agent receives a feedback from the environment that measures how good/bad was the chosen action; the so called *reward*. The MAB strategy used here is the Upper Confidence Bound (UCB) strategy where uncertainty in the action-value estimates is used for balancing exploration and exploitation. In our application, we assume the BS as the agent and we define the actions and reward as follows:

- **Actions**: An action, a, is defined as the choice of a pair composed of a DL UE and UL UE. The action at TTI t, $A\left(t\right)$, is chosen from the action set, $\mathcal{A}$, that is composed by all DL UE and UL UE pairs, i.e., $a \in \mathcal{A} = \mathcal{I} \times \mathcal{J}$ where the operator $\times$ represents the Cartesian product.
- **Reward**: The reward corresponding to the action taken at TTI t, $R\left(t\right)$, is given by:

$$R\left(t\right) = \frac{r_{i,j}^{\mathrm{tot,d}}\left(t\right) + r_{i,j}^{\mathrm{tot,u}}\left(t\right)}{2 \cdot F \cdot w_M}. \quad (5)$$

In our model, $R\left(t\right)$ is a number between 0 and 1. The reward is 0 when the previous transmission in uplink and downlink had an effective data rate equal to zero, i.e., the packets could not be decoded[2]. On the other hand, the reward is 1 when both link directions transmitted at the maximum possible data rate, i.e., $F \cdot w_M$ in uplink and $F \cdot w_M$ in downlink.

In order to estimate the expected reward for each action $a \in \mathcal{A}$ at TTI t, $\overline{R}\left(a, t\right)$, we use the average of the rewards yield per each action as in the following expression:

$$\overline{R}(a, t) = \frac{N\left(t-1, a\right) \overline{R}(a, t-1)}{N\left(t, a\right)} + \frac{\overline{R}(a, t)}{N\left(t, a\right)}, \quad (6)$$

[2]Note that in order to calculate $r_{i,j}^{\mathrm{tot,d}}\left(t\right)$ and $r_{i,j}^{\mathrm{tot,u}}\left(t\right)$ in the reward, we consider the actual inter UE channel since this is the effective received data rate in donwlink and uplink, respectively. In downlink, this can be estimated by means of Hybrid Automatic Repeat Request (HARQ) feedback (ACKs/NACKs).

Algorithm 1: UCB algorithm

for $t = 1; t \leq T; t = t + 1$ **do**
  Choose action $A(t)$ according to Eq. (7);
  for $f = 1; f \leq F; f = f + 1$ **do**
    Estimate SINR using the available CSI
    Select MCS according to the method described in section III
  end
  Transmits data with the configured MCS in uplink/donwlink
  Based on the received data rate, calculate reward $R(t)$ according to Eq. (5)
  Update average reward $\overline{R}$ according to Eq. (6)
end

TABLE I: Simulation Parameters

Parameters	Value
Cell Ring Area	$r_{min} = 30m$, $r_{max} = 100m$
Number of UEs	10
Monte Carlo Iterations	20
Carrier frequency	2.5 GHz
System bandwidth	180 kHz
Number of RBs	15
Number of subcarriers in an RB	12
Number of OFDM symbols in an RB	14
LOS path-loss model	$34.96 + 22.7 \log_{10}(d)$
NLOS path-loss model	$33.36 + 38.35 \log_{10}(d)$
LOS probability	$\min(\frac{18}{d}, 1) \cdot (1 - e^{\frac{-d}{36}}) + e^{\frac{-d}{36}}$
Shadowing st. dev. LOS	3 dB
Shadowing st. dev. NLOS	4 dB
Thermal noise power σ^2	-116.4 dBm/channel
Average user speed	3 km/h
User antenna height	1.5 m
BS antenna height	10m
SI cancelling level [β]	-110 dB
UCB constant c	5
Forgetfulness factor τ	0.9

where $N(t, a)$ consists in the number of times that action a was chosen until TTI t.

The choice of the action at TTI t, $A(t)$, is as follows [8]:

$$A(t) = \arg\max_{a \in \mathcal{A}} \left[\overline{R}(a, t) + c \cdot \sqrt{\frac{\ln(t)}{N(t, a)}} \right]. \tag{7}$$

In Algorithm 1 we summarize the steps of our proposed MAB-based solution for UE pairing.

V. NUMERICAL RESULTS

We implemented the main aspects of the system model presented in section III in a computational simulator in Python. The main parameters of our model are presented in Table I and are in accordance with [11]. We assume that the BS is placed at coordinate (0, 0) and UEs are uniformly distributed across a circular ring where the smaller radius is 30 m and the larger radius is 100 m.

The performance of our proposed algorithm is compared with the following benchmark solutions:

- Random (RND): At each TTI, UE pairs are chosen at random;
- Proportional Fair (PF): At each TTI, the chosen pair is the one that provides the highest value of F, as described by $F = \left(r_{i,j}^{\text{tot,d}}(t) + r_{i,j}^{\text{tot,u}}(t)\right) / \left(\overline{r}_i^{\text{tot,d}}(t) + \overline{r}_j^{\text{tot,u}}(t)\right)$ where $\overline{r}_i^{\text{tot,d}}(t)$ and $\overline{r}_j^{\text{tot,u}}(t)$ are, respectively, defined by

$$\overline{r}_i^{\text{tot,d}}(t) = \tau \cdot \overline{r}_i^{\text{tot,d}}(t-1)) + (1-\tau) \cdot r_{i,j}^{\text{tot,d}}(t) \tag{8}$$

$$\overline{r}_j^{\text{tot,u}}(t) = \tau \cdot \overline{r}_j^{\text{tot,u}}(t-1) + (1-\tau) \cdot r_{i,j}^{\text{tot,u}}(t); \tag{9}$$

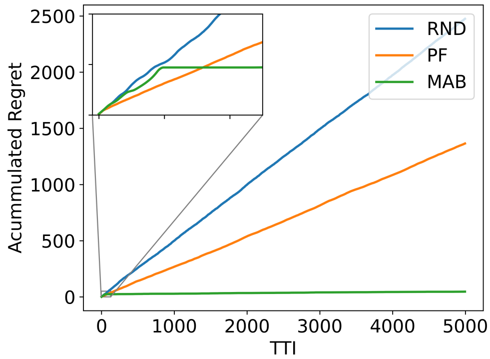


Fig. 2: Acummulated regret of different solutions.

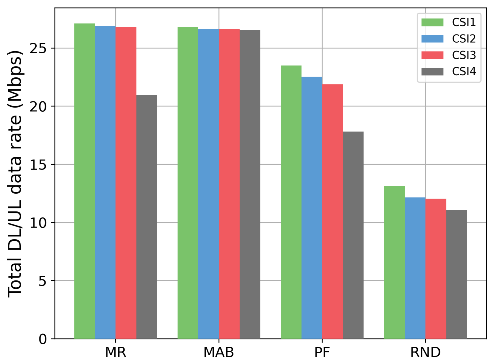


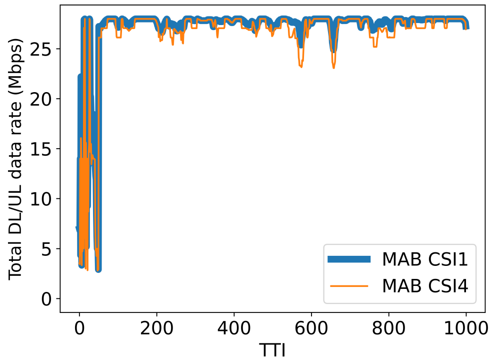
Fig. 3: Average total data rate UL/DL under different scenarios of CSI

- Maximum Rate (MR): At each TTI, the chosen pair is the one that provides the maximum sum of estimated transmit data rate in both uplink and downlink. The transmit data rate depends on the available CSI as explained in section III;

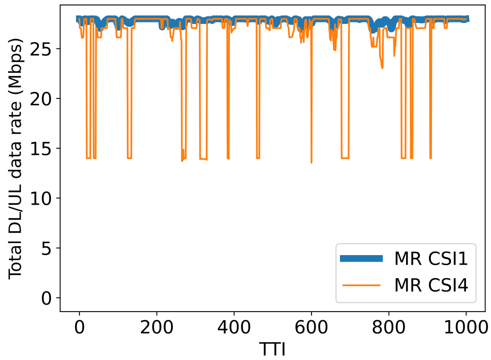
The performance of the algorithms can be measured in terms of regret. Regret is defined as the difference between the reward of the best possible pair and the reward of the pair selected by the algorithm. Fig. 2 shows the regret of each algorithm where we are using the result of maximum rate algorithm for the upper limit of data throughput for each TTI. During the first iterations the proposed algorithm could identify a good strategy as shown by the value of regret during the transmission.

In Fig. 3 we calculate the average total data rate for each algorithm. As we can see, the CSI degradation from CSI1 to CSI4 deteriorates the performance of all algorithms. However, our proposed solution is more robust to CSI uncertainty. For CSI1, CSI2, CSI3 and CSI4, UCB achieves 98.92%, 98.21% 98.19% and 97.91% of the maximum achievable data rate, respectively[3]. Interestingly, for CSI4, the UCB algorithm has an overall better performance with a gain of 26.44% over maximum rate solution with the same CSI conditions. Maximum rate presented a performance loss in total data rate of 22.57% from CSI1 to CSI4. As PF and RND are fairness-oriented algorithms, their total data rate are lower than MAB and maximum rate solutions.

[3]Maximum data rate is obtained from the use of MR solution with CSI1



(a) MAB



(b) MR

Fig. 4: Total UL/DL data rate per TTI for MAB and MR solutions.

Finally, Fig. 4 shows in details the CSI degradation effects in MAB and MR data rates along the TTIs. In Fig. 4a we can see an exploration period in the first TTIs but, after that, MAB CSI4 presents a very small performance loss compared to MAB CSI1. In contrast, Fig. 4b shows that the MR algorithm present a strong data rate degradation along the TTIs when compared to MR CSI1.

In summary, the great advantage of our solution is that it is more robust to the absence of inter-UE channel information. Our solution is capable of learning the best pairs by an intelligent exploration of the available actions, i.e., evaluating the impact on the effective received data rate of each UE pair. Although all algorithms experience performance degradation from CSI1 to CSI4 since link adaptation is performed with an estimated value for inter-UE channel gain, the algorithms MR and PF are more dependent on this since the UE priorities are based on estimated transmit data rates. Last but not least, we should emphasize that in MR and PF, in every TTI, the priority for each possible pair must be calculated, which increases computational complexity. In our proposed solution, only the expected reward for the last action should be updated and the action with maximum expected reward is chosen.

VI. Conclusion

In this article we propose the use of an ML solution based on reinforcement learning to solve the UE paring problem in point-to-multipoint FD systems. By using the MAB framework and employing UCB solution, simulation results have shown that our proposed method achieves quasi-optimum performance after the learning period when full CSI is available. Furthermore, when the CSI knowledge is degraded, our solution is more robust than benchmark solutions. Therefore, our proposed algorithm is capable to solve the UE pairing problem in FD networks without the need for inter-UE CSI knowledge.

References

[1] Ericsson. *Ericsson Mobility Report*. Tech. rep. June 2020.

[2] M. Amjad, F. Akhtar, M. H. Rehmani, M. Reisslein, and T. Umer. "Full-Duplex Communication in Cognitive Radio Networks: A Survey". In: *IEEE Communications Surveys Tutorials* 19.4 (Oct. 2017), pp. 2158–2191. ISSN: 2373-745X.

[3] X. Zhou, M. Sun, G. Y. Li, and B. Fred Juang. "Intelligent wireless communications enabled by cognitive radio and machine learning". In: *China Communications* 15.12 (Dec. 2018), pp. 16–48. ISSN: 1673-5447.

[4] Y. You, C. Qin, and Y. Gong. "Resource allocation for a full-duplex base station aided OFDMA system". In: *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. 2017, pp. 1–4.

[5] C. Nam, C. Joo, and S. Bahk. "Joint Subcarrier Assignment and Power Allocation in Full-Duplex OFDMA Networks". In: *IEEE Transactions on Wireless Communications* 14.6 (2015), pp. 3108–3119.

[6] T. T. Tran, V. N. Ha, L. B. Le, and A. Girard. "Dynamic Resource Allocation for Full-Duplex OFDMA Wireless Cellular Networks". In: *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*. 2016, pp. 1–5.

[7] R. Sultan, L. Song, K. G. Seddik, Y. Li, and Z. Han. "Mode selection, user pairing, subcarrier allocation and power control in full-duplex OFDMA HetNets". In: *2015 IEEE International Conference on Communication Workshop (ICCW)*. 2015, pp. 210–215.

[8] Richard Sutton. *Reinforcement learning : an introduction*. Cambridge, Massachusetts London, England: The MIT Press, 2018. ISBN: 978-0262039246.

[9] P. Pourbaba, S. Ali, K. B. S. Manosha, and N. Rajatheva. "Multi-Armed Bandit Learning for Full-Duplex UAV Relay Positioning for Vehicular Communications". In: *2019 16th International Symposium on Wireless Communication Systems (ISWCS)*. 2019, pp. 188–192.

[10] J. Tong, L. Fu, and Z. Han. "Throughput Enhancement of Full-Duplex CSMA Networks via Adversarial Multi-Player Multi-Armed Bandit". In: *2019 IEEE Global Communications Conference (GLOBECOM)*. 2019, pp. 1–6.

[11] P. Heino, Juha Meinilä, Pekka Kyösti, Lassi Hentila, Tommi Jämsä, Essi Suikkanen, E. Kunnari, and Milan Narandzic. "CP5-026 WINNER+ D5.3 v1.0 WINNER+ Final Channel Models". In: (Jan. 2010).