Reinforcement Learning Based Coexistence in Mixed 802.11ax and Legacy WLANs

Fabián Frommel*†, Germán Capdehourat*† and Federico Larroca†
*Ceibal, Uruguay. Email: {ffrommel,gcapdehourat}@ceibal.edu.uy
†Facultad de Ingeniería, Universidad de la República, Uruguay. Email: {fabian.frommel,gcapde,flarroca}@fing.edu.uy

***Abstract*—The new 802.11 amendment, 802.11ax, represents a significant shift in the WLAN operation, specially in the MAC layer where the access mechanism is now OFDMA. In particular, the Access Point (AP) is now responsible for scheduling the terminals' transmissions, which avoids collisions and results in an efficient usage of the spectrum. However, a full transition to this new technology is not foreseeable for several years, and until then mixed scenarios that also include legacy stations will be predominant. In this context, where both the AP and the legacy stations use CSMA/CA to access the channel, a very challenging aspect is the coexistence between both types of stations, where naturally the AP should have priority but legacy stations should not be excluded. In this paper we present a deep reinforcement learning system that adjusts the contention window so as to maximize a certain notion of fairness. Differently to previous proposals, none of which to the best of our knowledge focused on this mixed scenario, the choice of parameters that characterize the environment is informed on existing 802.11 models. This results for instance in a stable choice of the contention window and larger throughputs. Thorough simulations corroborate the performance of the proposed method, which we make available at https://github.com/ffrommel/RLinWiFi.**

***Index Terms*—CSMA/CA, OFDMA, Fairness, Deep Reinforcement Learning**

I. Introduction

IEEE 802.11 or Wi-Fi, as it is commercially known, is the most popular access technology. In terms of traffic, Wi-Fi networks have accounted for approximately 5 times more traffic than mobile networks in 2020 [1]. Even when considering handheld devices only, traffic accessed through Wi-Fi accounted for a portion of over 60%. Furthermore, the COVID pandemic has reverted a historical trend towards cellular access.

This popularity has resulted in extremely dense scenarios where, after several amendments that strived at increasing raw data rates, the traditional CSMA/CA access mechanism has been singled out as the system's bottleneck. In this context, the latest major amendment, 802.11ax or Wi-Fi 6 [2], represents a significant shift as it turned to OFDMA for improved efficiency. It is now the Access Point (AP) that schedules the terminals'[1] use of the spectrum [3].

However, a full transition to 802.11ax is expected to take several years. In the meantime, coexistence scenarios are to be the rule, where the legacy terminals compete for the medium with the AP, which should naturally have priority access as it schedules 802.11ax terminals. This poses the challenge of fair resource sharing between legacy and 802.11ax terminals. The typical approach to this kind of problems is to consider a model of the system and adjust operation parameters so as to optimize an objective function that in turn enforces a certain notion of fairness (see for instance [4]–[6]). However, to the best of our knowledge no satisfactory model for this coexistence mixed scenario exists to date.

In this paper we turn instead to reinforcement learning to solve this problem [7]. That is to say, we take a data-driven approach, where the AP learns to optimize the operation parameters based on its past experience. However, and differently to previous efforts in this direction [8], we design a model-informed learning system. That is to say, in the definition of the environment's state, we consider those indicators that are relevant according to existing models [9]. The actual mapping between these indicators and the reward, as well as choosing the next action, are left to the learning module.

In particular, we focus on adapting the contention window, as it is easily distributed from the AP to the terminals and has a large impact on the resulting resource distribution. As we will show through extensive simulations including several traffic patterns (UDP, TCP, uplink and/or downlink), this careful state definition avoids unstable choices of the contention window, resulting in a larger throughput in legacy-only scenarios than both a static choice or past proposals. As we mentioned before, the mixed scenario calls for the definition of a reward that takes into account fairness between legacy and 802.11ax terminals. The proposed system is flexible enough to accommodate this variant, where we show through simulations that our proposal is capable of choosing a contention window for each type of terminal that results in a fair distribution of throughput between them.

The rest of the article is structured as follows. After discussing previous efforts to include machine learning in general and reinforcement learning in particular to Wi-Fi networks in the next section, we present a first version of our proposal in Sec. III. A complete system that considers the coexistence scenario is presented and evaluated in Sec. IV, after which the article is concluded in Sec. V.

II. Reinforcement Learning in Wi-Fi

Many models have been developed for Wi-Fi networks, covering different cases such as saturated scenarios, UDP and TCP traffic, multirate WLANs, and qualitiy of service/experience (QoS/QoE) aspects [9]. All of them have been extremely

[1]We will use the term *terminal* or *station* interchangeably.

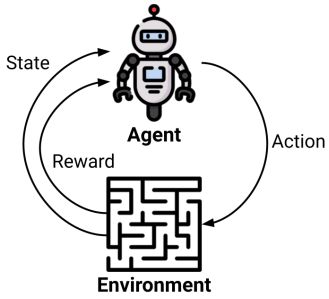

Fig. 1. Basic reinforcement learning model. The agent observes the environment and takes an action, which produces a reward. This is sequentially repeated and the objective is to maximize the long-term averaged reward.

useful for understanding the operation of the network, identifying problems and proposing new amendments to solve them. However, all of them have limitations that make their direct application in real-time decision making extremely difficult. The great variability that a Wi-Fi operational network poses in the real world turns out to be very complex for the development of mathematical models that cover all possible cases. In this context, it is precisely where the use of machine learning (ML) is most appropriate, since it enables to develop novel models from the data that represents each of the different scenarios and situations that could occur in real network operation [9].

Within ML, reinforcement learning (RL) is probably the best approach if what we are looking for is to solve sequential decision-making problems, such as optimizing (in some sense) the operation of the network. As we can see in Figure 1, the basic RL model has an agent that interacts with its environment to make decisions. The goal is to find a policy that determines the optimal actions to be taken according to the state of the system at each moment. To achieve this there are different methods, such as those based on the estimation of a value function (Q-learning) or actor-critic models (for more details refer to [10]).

With the rise of deep learning (DL) in recent years, RL methods have also integrated these techniques into the algorithms, such as the iconic autonomous Atari player developed based on deep reinforcement learning (DRL) [11]. In this case, a neural network is used to learn the Q-value function, the so-called Deep Q-Network (DQN), which is a suitable approach for high-dimensional continuous state spaces.

Many recent works focus on ML and RL algorithms applied to Wi-Fi networks [12], which hints at the current relevance of the research topic. Different applications and use cases are addressed with these techniques, such as link configuration, channel access, beamforming and spatial reuse, among many others (refer to [13] for more examples).

As previously mentioned, the present work is focused on the coexistence scenario that will dominate WLANs during the already ongoing transition from legacy Wi-Fi to 802.11ax. Although there are some previous works on coexistence (e.g. [14], [15]), all of them are related to the mixed scenarios between Wi-Fi and other technologies (e.g. LTE in unlicensed bands or cognitive networks), but none of them focuses on our case of interest.

III. A Model-Informed Learning System for Adjusting the Contention Window

A. Discussion

We propose an RL-based solution, which enables the Wi-Fi APs to dynamically select the optimal network parameters based on data collected from the network operation. Similarly to other previous works, the network parameter selected as optimization variable is the MAC layer maximum contention window (CW)[2], given its key role in the medium access control mechanism based on CSMA/CA.

However, there are some important contributions in our proposal that we highlight next. First off, our main requirement is that we are looking for a solution that is feasible to implement. To this end, the proposed system needs the participation of the AP only, unlike other proposals where terminals are also involved in the RL algorithm loop or have to modify the standard backoff mechanism of CSMA/CA (e.g. [16], [17]).

We base our model on the network visibility that the AP has, strengthening the central role for medium access control asigned in the new 802.11ax OFDMA scheme. Our approach follows the same direction as [8], where the authors present a centralized contention window optimization with DRL (CCOD). This method is based on deep deterministic policy gradient (DDPG), a DRL technique that runs in the network AP to dynamically select the optimal CW. Within the same general scheme, we introduce two major improvements to enhance the system performance.

On the one hand, based on classical 802.11 MAC layer analytical models, we consider an extended system space-state definition, which not only takes into account the network collision probability, but also the number of active terminals in the network. We will show its effect in the proposed algorithm performance, which reaches better and more robust results. On the other hand, we modify the reward function used in the RL framework. Utility functions such as the total network throughput are prone to the well known starvation issue, which could lead to unfair situations between different terminals. This point is crucial in a mixed scenario consisting of 802.11ax and legacy terminals. We address this issue by means of utility functions which lead to proportional fairness optimum solutions, presenting how to fit them into the RL scheme to be used.

Finally, and as we mentioned before, none of the previous works address the 802.11ax transition scenario. In this networks we have two different kind of stations, concerning the novel OFDMA medium access. While the 802.11ax capable terminals will follow the AP directives to know when to transmit, using the assigned spectrum portion (termed resource

[2]Please recall that this means that before transmitting, each station will uniformly at random select the number of slots it will spend in backoff from the interval $[0, \mathrm{CW}]$.

units or RUs) on each access round, legacy terminals will continue using CSMA/CA based access. The next section will present how the proposed algorithm deals with this situation in the RL framework defined.

B. Legacy-Only Scenario and the Environment's Definition

In this subsection we discuss the importance of an informed definition of the parameters that characterize the environment when using a RL algorithm. To this end, we will take as a baseline CCOD, the proposal we mentioned before [8], corresponding to a DRL algorithm that runs in the network AP to dynamically adjust the MAC layer CW in order to maximize the total throughput. In this case the system state model is the estimated collision probability in the network. As we will show shortly, it is not possible to appropriately control the system operation by means of this variable only.

Consider a scenario where an increasing number of legacy stations start transmitting uplink UDP traffic. This simple example may be analytically studied by means of classical 802.11 MAC layer mathematical models (such as Bianchi's seminal work [18]), from which an expression for the collision probability and total throughput may be obtained as a function of the number of transmitting stations and the configured CW. A simple grid search is enough to estimate the optimal CW, resulting in an increasing optimal CW as the number of stations increases, which in turn results in a collision probability that is roughly constant independently of the number of stations.

The previous discussion means that representing the environment's state with the collision probability only leads to ambiguous situations. Even when exposed to scenarios with different number of stations, the agent will not learn to estimate the optimal CW if using the collision probability only.

This is verified in the simulation result shown in Figure 2,[3] which compares the CW obtained by CCOD with the optimal value as estimated by the analytical model as the number of stations increases. In particular, five stations are initially transmitting, and a single new station is activated every 300 seconds until 25 are transmitting simultaneously. CCOD was trained under this same scenario, and the figure shows the results of the trained system. As we explained before, CCOD presents an unstable behavior when varying the number of terminals and the selected CW has large oscillations.

In order to address this model limitation, we propose to incorporate to the system space-state definition the number of active stations in the network as another state variable. It is worth noting that this number could be different to the total number of stations associated to the AP, as only *active* stations should be considered. That is to say only those stations that are actually contending for the medium access in a given time.

In our system, the RL-agent runs in the network AP, so both observation variables should be estimated by it. For this purpose, we compute the collision probability as in [8], taking

[3]All of the simulations we present here are based on ns-3 [19], PyTorch [20] and ns3-gym [21], using similar settings as in [8] (e.g. single AP, 20 MHz channel, single-user transmissions, frame aggregation disabled). Details are available at https://github.com/ffrommel/RLinWiFi.

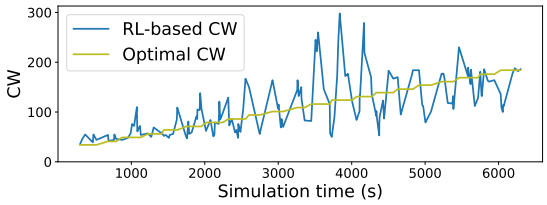

Fig. 2. The CW chosen by CCOD [8] as the number of stations increases from 5 to 25 (compared to the optimal choice computed based on an analytical model). Since CCOD represents the environment only through the collision probability, the chosen CW oscillates.

the mean and variance values for the latest time windows of the considered network history. To keep track of the active stations, the AP accounts only those stations that transmit a number of frames above a certain threshold, considering the same history length than for the collision probability estimation.

The new system state proposal results in a novel model-informed learning agent, that we will now show is able to adapt the CW properly to different changes in network conditions. In the experiments of the next subsections we will still use as reward function the total network throughput, to be able to compare the results in [8]. Then, in Section IV we will discuss about fairness issues and the coexistence scenarios between 802.11ax and legacy devices, and present a solution that fits into the same RL scheme.

1) Variations in the Number of Stations: The first case scenario analyzed corresponds to the same presented in Figure 2. Recall that the number of active stations in the network varies from 5 to 25 stations, all of them transmitting uplink UDP traffic to the AP. Similarly to CCOD, in order to train the RL agent, the exploration training phase covers all the different number of transmitting stations and also several CW choices. For this purpose a random noise is used at the agent in order to densely traverse the action-space. Finally, in the evaluation phase the noise is turned off, and the agent selects the CW at each stage according to the trained model.

Figure 3 presents the results obtained for the model-informed agent, showing how the selected CW evolves during the evaluation phase, as well as the CW corresponding to the optimal as computed by the 802.11 saturation mathematical model [18]. It is important to note that the agent is able to follow the network dynamics properly without oscillations. These results demonstrate that integrating the number of active stations into space-state model is a suitable approach, enabling the agent to accurately identify the system dynamics and adapt the CW accordingly.

2) Variations in Carried Traffic: Another relevant case scenario which is expected that the agent would be able to deal with is when the type of traffic changes during network operation. That is to say, what happens when traffic varies from UDP to TCP and viceversa. In this case it is relevant again to account for the number of active stations in the network, because when we have TCP traffic it is well known that only

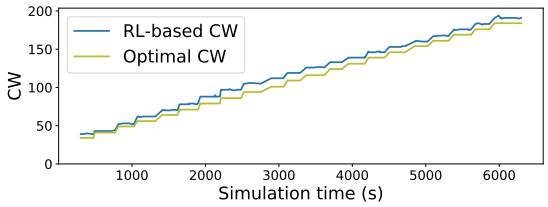

Fig. 3. The CW chosen by our model-informed learning agent as the number of stations increases from 5 to 25 (compared to the optimal choice computed based on an analytical model). The inclusion of the number of active stations to represent the environment results in the optimum CW being chosen.

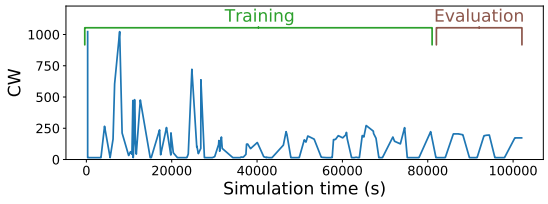

Fig. 4. Model-informed learning agent results when traffic type changes (UDP/TCP) during network operation. The definition of *active* stations in the environment characterization is key in choosing a smaller CW for TCP than UDP traffic.

a few stations are active each time (as most of them are waiting for TCP ACKs from the AP) [9]. This fact implies that when the total number of stations in the network increases, the optimal CW value for UDP traffic tend to be significantly higher than the corresponding one for TCP traffic.

For this purpose, a dynamic scenario was simulated, consisting of a constant number of 25 stations, which continuously sends traffic to the AP, but switching between UDP and TCP traffic. In this context, the RL agent was trained the same way than in the previous case, with a exploration phase covering different network traffic for the several possible values of the action-space (i.e. the CW values).

In Figure 4 we can see the evolution of the selected CW. As indicated in the graph, the first part of the simulation corresponds to the exploration phase, where the agent is being trained. As soon as this phase ends, the already trained agent starts its operation, where it reaches a stable result. Note that the assigned CW value is higher for UDP traffic than for TCP traffic.

The results shown in Figure 4 confirm that the agent is being able to converge properly in the learning phase, achieving a robust CW adaptation after completing this stage. In Table I the convergence results are detailed, confirming that the CW values selected by the agent for each type of traffic are in line with the theoretical optimum expected values.

IV. Coexistence Between IEEE 802.11ax and Legacy Terminals

A. Maximum Efficiency

We will now consider a scenario where an 802.11ax AP serves both 802.11ax and legacy terminals. To this end, let us first very briefly recall how the new standard operates in this

TABLE I
Model-informed learning agent results during evaluation using different type of traffic scenarios.

Traffic	n	CW	p	Throughput
TCP DL	25	15	0.06	31.3 Mbps
UDP UL	25	170	0.2	38.3 Mbps

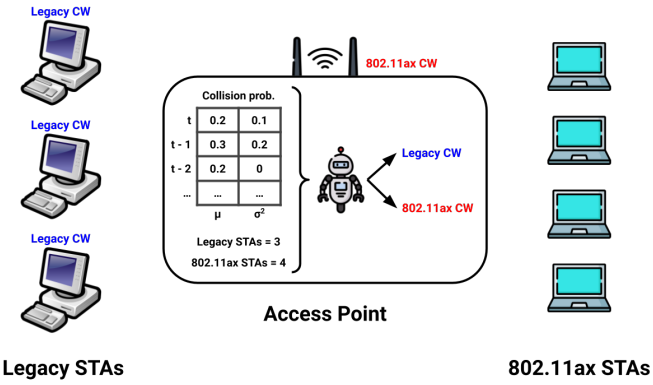

Fig. 5. Schematic representation and update process of CWs in the model-informed learning method. The agent takes into account a sequence of mean and variance collision probability, as well as the number of active legacy and ax stations (STAs) to choose the next legacy and ax CW.

scenario. In the uplink sense, 802.11ax terminals only transmit when the AP indicates so, which is performed through a so-called *Trigger Frame*. This special frame includes the list of stations that will transmit, as well as common and specific parameters to be used during the transmission (e.g. guard interval and RUs of each station respectively). This frame is immediately followed by the transmission of the indicated terminals.

So as to allow legacy terminals to operate, the AP still runs CSMA/CA to access the channel, both for trigger frames and downlink data (where each frame now includes transmissions for several 802.11ax terminals by means of OFDMA). Naturally, the AP should have preferential access as it serves several stations, typically by means of a smaller contention window. However, the precise contention window value to choose is not specified in the standard, and in this section we propose to set it dynamically through DRL.

Note that we now have two contention windows to set: one corresponding to the AP (which will affect transmissions from and to 802.11ax terminals) and another value corresponding to legacy terminals. It was necessary then to extend the system we proposed in the previous section to use an environment with terminals of these two types and an agent that, from the observation of the environment, can select the value of two CWs. In particular, the environment is now characterized by the collision probability as before, but we now include the number of both legacy and 802.11ax active stations. Figure 5 provides a schematic representation of the learning system we propose for mixed scenarios.

Let us illustrate the new scenario with an example. Consider

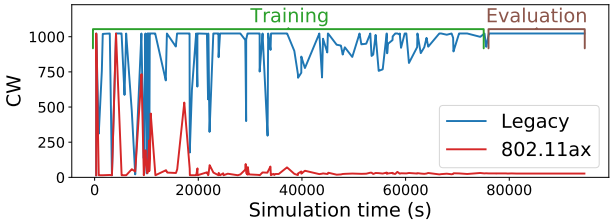

Fig. 6. Results for training the agent to maximize the total throughput when 10 legacy and ax terminals are transmitting uplink UDP traffic. The very large CW for legacy devices results in a total throughput of about 88.5 Mbps, where only about 3.5 Mbps are for the legacy devices.

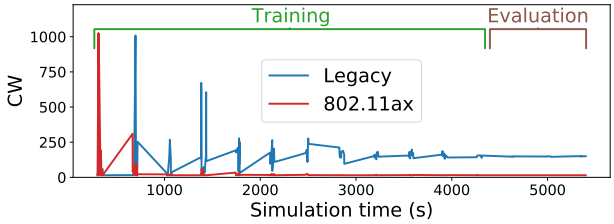

Fig. 7. Results for training the agent to maximize the total utility (Eq. (1)) when 10 legacy and ax terminals are transmitting uplink UDP traffic. Although ax terminals are still prioritized, legacy terminals obtain a much smaller CW (cf. Fig. 6).

UDP uplink transmissions, in a static scenario with a fixed number of $l = 10$ 802.11ax and $m = 10$ legacy stations. If we were to reward the system in terms of the total throughput only (as before), the obtained results are shown in Figure 6. Note that, as expected, the learning system prioritizes transmissions from 802.11ax terminals, given how their transmissions are much more efficient in the channel usage. In terms of throughput, the result is almost the complete exclusion of legacy terminals from transmissions: legacy stations registered a average value of only 3.5 Mbps, while 802.11ax averaged 85 Mbps. The challenge is thus to bring certain notion of fairness to the learning system, so as to avoid this undesirable behavior.

B. Fair Coexistence

The idea of fairness between terminals of different types in IEEE 802.11 networks has been studied mostly in the context of different data rates. In particular, the problem is posed in term of airtime fairness, which translates into a fair distribution of channel usage over time. There are variations on its implementation, such as token-bucket regulator [22], airtime deficit round-robin scheduler [23], contention window controller [24], and 802.11e TXOP [25]. Other works, however, use the notion of utility maximization to allow a balance between fairness and spectrum efficiency. Basically, the economics-inspired idea is that each user will get a certain increasing and concave utility out of the received throughput, and the objective is to maximize the overall welfare measured through the sum of these utilities. In the context of networking, it was first applied to congestion control in the seminal paper by Kelly et al. [26], but has been since extended to several other areas, including naturally Wi-Fi [27]. For instance, in [28] a modified CSMA algorithm is introduced that implements this idea and compares its performance with other fairness criteria, including airtime fairness.

This framework is particularly suited for (D)RL, as it is posed as an optimization problem. In particular, we consider a certain time window (e.g. the last second) and measure S_i, the throughput of each active station during that time, both for 802.11ax and legacy stations (i.e. $i = 1, \ldots, l + m$). We may thus define the reward as:

$$r = \sum_{i=0}^{l+m} \frac{1}{1-\alpha} S_i^{1-\alpha}, \tag{1}$$

TABLE II
RESULTS FOR THE CWS CHOSEN BY THE LEARNING ALGORITHM (IN BOLD), AND BY MANUALLY CHOOSING OTHER VALUES.

CW		Throughput [Mbps]			Utility
Legacy	ax	Legacy	ax	Total	Relative
150	**15**	**26.8**	**25.8**	**52.6**	**1.00**
75	75	28.4	20.7	49.1	0.91
300	15	23.9	25.7	49.6	0.94
75	15	27.6	22.4	50.0	0.94
300	75	24.3	24.5	48.8	0.92

which results in the so-called α-fairness [29]. The parameter α controls the balance between efficiency ($\alpha = 0$ maximizes the sum, as before), and complete fairness ($\alpha \to \infty$ results in max-min fairness, maximizing the rate of the station with the smallest rate). The so-called proportional fairness is a midpoint obtained by using $\alpha \to 1$, and the one we use in the example that follows, although the network administrator is free to choose any other utility function.

Figure 7 shows the results corresponding to the same scenario as before, but using this fairness-aware reward. It can be seen how now the legacy CW converges to a much smaller value. This results in a throughput for legacy and 802.11ax terminals of approximately 25 Mbps each, totaling about 50 Mbps. Although the total throughput is smaller than before, this lower efficiency allows for legacy stations to be able to use the network fairly.

As we mentioned before, no model exists for this mixed scenario. In order to verify that the agents obtains an optimal CW, Table II shows the resulting throughput and utility when manually changing the CWs. Note that the agent has chosen the minimum possible CW for the ax terminals (15). We thus evaluate what happens when we increase this value, whereas for the legacy ones we explore values around the chosen one. For instance, doubling the CW for legacy stations from 150 to 300 results in a lower throughput for them, whereas the ax stations are not able to increase their throughput, all in all resulting in a lower total utility.

V. CONCLUSIONS AND FUTURE WORK

Since the IEEE approved the new 802.11ax amendment, a new cycle of technological transition began for Wi-Fi networks. This is not new, as it happened before every time

a new amendment was approved. However, this transition to 802.11ax will be different from all previous ones, since for the first time the MAC layer access is modified. The traditional medium access control based on CSMA/CA, is replaced by OFDMA, giving to the AP all the resource allocation responsibility. This means that over the next years two very different access mechanisms will have to coexist in Wi-Fi networks, one for 802.11ax devices and the other for legacy ones.

Our work focuses precisely on this 802.11ax coexistence scenario, so relevant to the operation of Wi-Fi networks in the years to come. We propose an algorithm based on deep reinforcement learning (DRL), in order to properly solve resource allocation, by means of dynamically adapting the contention window. The DRL-based agent runs in the network AP, which emphasizes the central role that the new standard already granted to it for medium access control. A salient feature of our proposal is that it is a model-informed learning agent, as the state-space of the RL model is based on the estimation of the two main variables in previous analytical models for the MAC layer: collision probability and the number of active terminals in the network.

Extensive simulations show that the agent operates as expected in different scenarios, reacting properly when network conditions change, such as the number of terminals as well as in the traffic type. This results verify a robust behaviour of the agent, with stable choices of the CW and larger throughputs. Moreover, we further discuss on fairness issues, integrating a different reward function to the same RL scheme, in order to guarantee a fair traffic share between 802.11ax and legacy stations. In future work, this point could be further studied, for example evaluating parametric reward functions, which would enable network administrators to select the desired traffic sharing between 802.11ax and legacy terminals when configuring the APs.

REFERENCES

[1] Wireless network data traffic: worldwide trends and forecasts 2021–2026. Analysys Mason. [Online]. Available: https://www.analysysmason.com/research/content/regional-forecasts-/wireless-traffic-forecast-rdnt0/

[2] IEEE 802.11ax-2021 - IEEE Standard for Information Technology. https://standards.ieee.org/standard/802_11ax-2021.html.

[3] E. Khorov, A. Kiryanov, A. Lyakhov, and G. Bianchi, "A Tutorial on IEEE 802.11ax High Efficiency WLANs," *IEEE Communications Surveys Tutorials*, vol. 21, no. 1, pp. 197–216, 2019.

[4] P. Gallo, K. Kosek-Szott, S. Szott, and I. Tinnirello, "CADWAN: A Control Architecture for Dense WiFi Access Networks," *IEEE Communications Magazine*, vol. 56, no. 1, pp. 194–201, 2018.

[5] D.-J. Deng, C.-H. Ke, H.-H. Chen, and Y.-M. Huang, "Contention window optimization for ieee 802.11 DCF access control," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5129–5135, 2008.

[6] K. Hong, S. Lee, K. Kim, and Y. Kim, "Channel condition based contention window adaptation in ieee 802.11 wlans," *IEEE Transactions on Communications*, vol. 60, no. 2, pp. 469–478, 2012.

[7] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep Reinforcement Learning: A Brief Survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[8] W. Wydmański and S. Szott, "Contention Window Optimization in IEEE 802.11ax Networks with Deep Reinforcement Learning," in *2021 IEEE Wireless Communications and Networking Conference (WCNC)*, 2021, pp. 1–6.

[9] F. Larroca and F. Rodríguez, "An Overview of WLAN Performance, Some Important Case-Scenarios and Their Associated Models," *Wirel. Pers. Commun.*, vol. 79, no. 1, p. 131–184, Nov 2014. [Online]. Available: https://doi.org/10.1007/s11277-014-1846-4

[10] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html

[11] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," 2013, nIPS Deep Learning Workshop 2013. [Online]. Available: http://arxiv.org/abs/1312.5602

[12] S. Szott, K. Kosek-Szott, P. Gawłowicz, J. T. Gómez, B. Bellalta, A. Zubow, and F. Dressler, "Wi-fi meets ml: A survey on improving ieee 802.11 performance with machine learning," *IEEE Communications Surveys Tutorials*, vol. 24, no. 3, pp. 1843–1893, 2022.

[13] F. Wilhelmi, S. Barrachina-Munoz, B. Bellalta, C. Cano, A. Jonsson, and V. Ram, "A flexible machine-learning-aware architecture for future wlans," *IEEE Communications Magazine*, vol. 58, no. 3, pp. 25–31, 2020.

[14] S. Mosleh, Y. Ma, J. D. Rezac, and J. B. Coder, "Dynamic spectrum access with reinforcement learning for unlicensed access in 5g and beyond," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 2020, pp. 1–7.

[15] A. H. Y. Abyaneh, M. Hirzallah, and M. Krunz, "Intelligent-cw: Ai-based framework for controlling contention window in wlans," in *2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, 2019, pp. 1–10.

[16] L. Zhang, H. Yin, Z. Zhou, S. Roy, and Y. Sun, "Enhancing wifi multiple access performance with federated deep reinforcement learning," in *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*, 2020, pp. 1–6.

[17] R. Ali, N. Shahin, Y. B. Zikria, B.-S. Kim, and S. W. Kim, "Deep reinforcement learning paradigm for performance optimization of channel observation–based mac protocols in dense wlans," *IEEE Access*, vol. 7, pp. 3500–3511, 2019.

[18] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, 2000.

[19] ns-3. https://www.nsnam.org/.

[20] PyTorch. https://pytorch.org/.

[21] ns3-gym: OpenAI Gym integration. https://apps.nsnam.org/app/ns3-gym/.

[22] H. SHI, R. V. Prasad, E. Onur, and I. Niemegeers, "Fairness in Wireless Networks:Issues, Measures and Challenges," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 5–24, 2014.

[23] R. Riggio, D. Miorandi, and I. Chlamtac, "Airtime Deficit Round Robin (ADRR) packet scheduling algorithm," in *2008 5th IEEE International Conference on Mobile Ad Hoc and Sensor Systems*, 2008, pp. 647–652.

[24] T. Joshi, A. Mukherjee, Y. Yoo, and D. P. Agrawal, "Airtime Fairness for IEEE 802.11 Multirate Networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 4, pp. 513–527, 2008.

[25] L. B. Jiang and S. C. Liew, "Proportional fairness in wireless LANs and ad hoc networks," in *IEEE Wireless Communications and Networking Conference, 2005*, vol. 3, 2005, pp. 1551–1556 Vol. 3.

[26] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research society*, vol. 49, no. 3, pp. 237–252, 1998.

[27] A. Ferragut and F. Paganini, "Resource allocation over multirate wireless networks: A network utility maximization perspective," *Computer Networks*, vol. 55, no. 11, pp. 2658–2674, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128611001824

[28] S. Krishnan and P. Chaporkar, "Stochastic approximation based on-line algorithm for fairness in multi-rate wireless lans," *Wireless Networks*, vol. 23, no. 5, pp. 1563–1574, Jul 2017. [Online]. Available: https://doi.org/10.1007/s11276-016-1243-x

[29] T. Lan, D. Kao, M. Chiang, and A. Sabharwal, "An axiomatic theory of fairness in network resource allocation," in *2010 Proceedings IEEE INFOCOM*, 2010, pp. 1–9.