Decentralized Spatial Reuse Optimization in Wi-Fi: An Internal Regret Minimization Approach

Francesc Wilhelmi
Wireless Networking
Universitat Pompeu Fabra, Spain
Email: francisco.wilhelmi@upf.edu

Boris Bellalta
Wireless Networking
Universitat Pompeu Fabra, Spain
Email: boris.bellalta@upf.edu

Miguel Casasnovas
Wireless Networking
Universitat Pompeu Fabra, Spain
Email: miguel.casasnovas@upf.edu

Aleksandra Kijanka
Wireless Networking
Universitat Pompeu Fabra, Spain
Email: aleksandra.kijanka@upf.edu

Miguel Calvo-Fullana
Wireless & Secure Communications
Universitat Pompeu Fabra, Spain
Email: miguel.calvo@upf.edu

***Abstract*—Spatial Reuse (SR) is a cost-effective technique for improving spectral efficiency in dense IEEE 802.11 deployments by enabling simultaneous transmissions. However, the decentralized optimization of SR parameters—transmission power and Carrier Sensing Threshold (CST)—across different Basic Service Sets (BSSs) is challenging due to the lack of global state information. In addition, the concurrent operation of multiple agents creates a highly non-stationary environment, often resulting in suboptimal global configurations (e.g., using the maximum possible transmission power by default). To overcome these limitations, this paper introduces a decentralized learning algorithm based on regret-matching, grounded in internal regret minimization. Unlike standard decentralized "selfish" approaches that often converge to inefficient Nash Equilibria (NE), internal regret minimization guides competing agents toward Correlated Equilibria (CE), effectively mimicking coordination without explicit communication. Through simulation results, we showcase the superiority of our proposed approach and its ability to reach near-optimal global performance. These results confirm the not-yet-unleashed potential of scalable decentralized solutions and question the need for the heavy signaling overheads and architectural complexity associated with emerging centralized solutions like Multi-Access Point Coordination (MAPC).**

I. Introduction

The proliferation of wireless devices and the exponential growth in traffic demands have pushed Wi-Fi to undergo an ambitious transformation and adopt significant architectural changes [1]. IEEE 802.11 relies on Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA), which has been proven effective in sparse networks, but performs poorly in dense deployments where neighboring Basic Service Sets (BSSs) share time resources inefficiently (e.g., by imposing conservative carrier sensing policies that lead to excessive contention). To address this, among many other features, recent amendments such as 802.11ax (Wi-Fi 6) and 802.11be (Wi-Fi 7) have introduced Spatial Reuse (SR), which allows devices to ignore inter-BSS interference below a certain Carrier Sense Threshold (CST) (technically referred to as Overlapping Basic Service Set/Packet Detect (OBSS/PD) threshold), thereby enabling concurrent transmissions on the same frequency channels. By combining CST adjustment and Transmit Power Control (TPC), SR can potentially unlock more simultaneous transmissions and thus increase network performance.

However, Wi-Fi's decentralized nature (i.e., BSSs operate autonomously, lacking information regarding the channel state, traffic load, or configuration of neighboring networks), hinders the optimization of SR. When multiple uncoordinated BSSs attempt to concurrently optimize CST and transmit power—whose relationship is highly non-linear and coupled—, they create a non-stationary environment, where the interactions between devices change constantly. Under these conditions, online learning approaches emerge as a compelling solution to achieve global SR configurations [2]. Still, state-of-the-art decentralized learning algorithms, which typically seek to maximize individual utility (external regret minimization), may fail to find globally optimal and stable solutions, thus leading to inefficient Nash Equilibrium (NE) [3].

Currently, the Task Group bn (TGbn) is developing the next generation of the 802.11 standard, 802.11bn (Wi-Fi 8) [1], [4], where Multi-Access Point Coordination (MAPC) is proposed to address some of the decentralization challenges. While coordination can enable optimal configurations, it introduces significant scalability barriers, including substantial signaling overhead and synchronization requirements. This is the reason why, so far, MAPC limits coordination to a maximum of two Access Points (APs). In this paper, we propose a decentralized, online learning alternative that achieves the performance benefits of coordination without the associated overhead. We introduce an algorithm based on regret-matching [5] that aims to minimize the internal regret of agents. Unlike standard algorithms that compare a strategy's performance against a fixed best alternative, internal regret minimization algorithms evaluate how much better an agent would have performed if it had swapped one specific action for another. By doing this, our algorithm guides independent BSSs toward a Correlated Equilibrium (CE)—a state of implicit coordination that maximizes global welfare. Our contribution demonstrates that scalability and optimality are not mutually exclusive in next-generation Wi-Fi networks. We show that by engineering the learning objective to minimize internal regret, Wi-Fi devices can learn efficient SR patterns, questioning the need for complex, centralized coordination architectures.

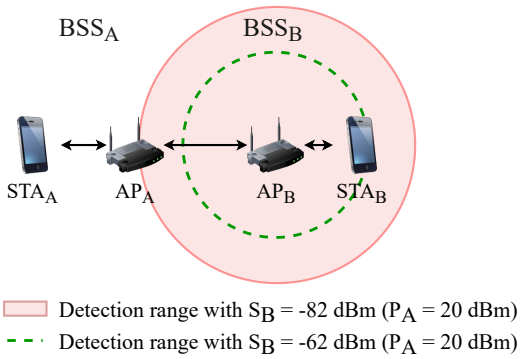


Figure 1: Two BSSs, each comprising one AP and one Station (STA), coexist within the same area. The sensitivity range from AP_B indicates whether the transmissions from AP_A are ignored (AP_A is outside the circle) or not (AP_A is inside the circle). P_X and S_Y denote the power and sensitivity used by AP X and Y, respectively.

The remainder of this paper is structured as follows: Section II reviews Wi-Fi's SR and the related literature. Section III formulates the SR problem as a multi-player game and describes the proposed internal regret minimization solution, which is subsequently evaluated in Section IV. Finally, Section V provides concluding remarks.

II. Learning Spatial Reuse in Wi-Fi

A. Spatial Reuse in Wireless Communications

In wireless communications, SR refers to the ability of wireless devices to perform simultaneous transmissions using the same or overlapping frequency channels. This can be done by properly tuning parameters such as the transmit power P (*determining the interference caused to others*) and CST S (*determining the interference tolerance from others*) [6]. In particular, the transmit power determines the reach of transmissions and the potential quality of the signal at the intended receiver (see Fig. 1). A *high power* (e.g. 20 dBm) would lead to a better Signal-to-Noise-Ratio (SNR), but would incur higher interference to other devices. A *low power* (e.g., 3 dBm) would create less interference to other devices, but results in a poorer SNR. On the other hand, adjusting the listening sensitivity, namely CST, in the context of Listen-Before-Talk (LBT) (e.g., used for preamble detection) dictates the device's aggressiveness in accessing the medium. Using a low threshold (e.g., -82 dBm) would lead to a higher number of transmission deferrals, while a high threshold (e.g., -62 dBm) would allow the device to transmit more often (but potentially under higher interference conditions).

B. Spatial Reuse in Wi-Fi

The implementation of SR in Wi-Fi has evolved since the 802.11ax amendment, which first introduced OBSS/PD-based SR [6]. This feature allows devices to ignore inter-BSS interference by using a CST threshold higher than the Clear Channel Assessment (CCA) for frames that carry a different BSS Color (i.e., inter-BSS transmissions). However, to protect any ongoing transmissions, a strict transmit power limitation is enforced during an SR Transmission Opportunity (TXOP) that has been obtained by using a higher CST. A few years later, IEEE 802.11be introduced Parametrized Spatial Reuse (PSR) [7], which facilitates concurrent (uplink) transmissions through TXOP sharing. Similar to OBSS/PD, PSR requires the secondary transmitter to adhere to power limits imposed by the primary AP (e.g., based on its tolerated interference). Looking ahead, 802.11bn aims to standardize Coordinated Spatial Reuse (Co-SR) [8], a natural extension of these mechanisms tailored to the MAPC framework.

C. Reinforcement Learning Solutions for Spatial Reuse

To address the SR problem, decentralized Reinforcement Learning (RL) has been widely adopted, including works that have applied Q-Learning and Multi-Armed Bandit (MAB) to optimize the transmit power and CST [3], [9]. In these works, RL algorithms are used to maximize individual utilities (e.g., per-BSS throughput), which proves effective for addressing the inherent complexity of the problem in certain scenarios. However, when applied concurrently by multiple agents, achieving optimal equilibria becomes difficult given the aggressive strategies learned by the agents in such a competitive environment. To address this, other solutions have been proposed, including reward-sharing bandits [8], cooperative bandits [10], mechanisms based on interferer identification [11], and complex wireless state characterization through Deep Reinforcement Learning (DRL) [12].

Non-coordinated strategies, in line with Wi-Fi's decentralized nature, largely rely on rational learning and partial information (e.g., external regret minimization), which trap agents in inefficient equilibria dominated by aggressive strategies (e.g., all the agents use a high transmit power as a protective measure). To overcome the limitations of decentralized learning, Game Theory stands as a promising tool to achieve the high efficiency of coordination without incurring large overheads. In this regard, we find early works proposing cooperative games (e.g., Shapley values, bargaining) in areas like cognitive radio or mesh networking [13], [14], focusing primarily on power control and channel allocation. Similarly, [15], [16] provided no-regret solutions for power control in wireless networks. In the context of Wi-Fi, there is a lack of literature applying internal regret minimization strategies, which is precisely the gap this paper aims to fill.

III. Problem Formulation and Solution Proposal

A. Formulation as an Online Decision-Making Game

We formulate the optimization of SR in Wi-Fi as a Multi-Agent (MA) MAB game, where a set of agents $\mathcal{N} = \{1, \dots, N\}$ operate independently and iteratively select an action from a set $\mathcal{A} = \{1, \dots, K\}$, leading to a joint action profile $\boldsymbol{a} = (a_1, \dots, a_N)$. The game is divided into T slots, where each slot represents a decision-making opportunity for the agents, and the goal of each agent is to optimize its configuration and maximize its own performance (e.g., throughput). In a given iteration $t \in \{1, \dots, T\}$, agents obtain bandit feedback, meaning that they only observe their own reward r_n, which depends on the joint strategy profile $r_n^t = f_r(a_n^t, a_{-n}^t)$, where $f_r(\cdot)$ is the reward function and a_{-n}^t denotes the actions selected by all other BSSs. In the SR problem, the action space

for each BSS consists of a discrete pair of transmit power and CST values $a = (P, S)$, where $P \in \mathcal{P}$ and $S \in \mathcal{S}$ (with $\mathcal{P}$ and $\mathcal{S}$ being the transmit power and sensitivity action spaces, respectively). The fundamental challenge in this problem is that r_n is non-convex and non-stationary as other BSSs adapt dynamically. Consequently, standard optimization approaches (e.g., external regret minimization) typically lead to NE, which are often inefficient in interference channels [17].

B. External vs. Internal Regret

In Wi-Fi, SR interactions are complex due to the consequent contention, collisions, and other adverse effects stemming from decentralization (e.g., unexpected collisions due to hidden nodes, additive interference). Moreover, the bandit (partial) feedback means agents only observe the payoff of the actions they play in each iteration. To define a strategy for individual agents, we can opt for minimizing either the *external* or the *internal* regret [18].

The external regret (R_n^{ext}) looks at the best-performing action in hindsight (i.e., the difference between the received reward and the reward that could have been achieved by playing the single best fixed action over the entire game) to quantify the performance of an action-selection algorithm. In particular, the external regret for BSS n at time horizon T is

$$R_n^{\text{ext}}(T) = \max_{k \in \mathcal{A}_n} \sum_{t=1}^{T} r_n(k, a_{-n}^t) - \sum_{t=1}^{T} r_n(a_n^t, a_{-n}^t), \qquad (1)$$

where the first term of the equation refers to the performance of the best fixed strategy, and the second term represents the actual obtained performance. Here, $r_n(k, a_{-n}^t)$ denotes the hypothetical reward the agent would have obtained if it had played action k in iteration t, given a_{-n}^t.

While external regret minimization works well in static setups, it is often insufficient for multi-agent settings, as it fails to capture the non-stationary dynamics of adaptive opponents (i.e., the external regret looks more at the history of the actions rather than at game dynamics). This typically drives the system to NE states characterized by aggressive, selfish behaviors (e.g., employ maximum transmit power) that result in suboptimal network-wide performance. To address this, the internal regret, instead of comparing actions alone, compares the loss of an online algorithm to the loss of a modified online algorithm, which consistently replaces one action with another. This favors the enforcement of CE, which, contrary to NE (which assumes independent strategies), allows correlating the agents' joint actions, implicitly capturing how the system responds to specific moves. This often leads to higher system-wide efficiency and fairness. In particular, the cumulative internal regret achieved by BSS n at time T is

$$R_n^{\text{int}}(T) = \max_{j,k \in \mathcal{A}_n} \sum_{t: a_n^t = j} \left(r_n(k, a_{-n}^t) - r_n(j, a_{-n}^t) \right), \qquad (2)$$

where the summation is done only for the iterations where action j was played (i.e., $t : a_n^t = j$), and $r_n(k, a_{-n}^t)$ is the hypothetical reward the agent would have obtained if it had played k instead of j in iteration t.

Algorithm 1 Regret-matching for decentralized spatial reuse

1: **Input:** $\mathcal{A}$ (including power/sensitivity pairs)
2: **Initialize:** $Q_{j \to k} \leftarrow 0, \forall j, k \in \mathcal{A}$, $\boldsymbol{\pi} \leftarrow \{\frac{1}{|\mathcal{A}|}, ..., \frac{1}{|\mathcal{A}|}\}$, $\mu \leftarrow 2(|\mathcal{A}|-1)$ (stickiness factor), $\lambda \leftarrow 0.95$ (decay factor)
3: **for** $t = 1, \ldots, T$ **do**
4: **// Action selection & reward estimation**
5: $a^t \leftarrow \arg\max_{k \in \mathcal{A}}, \boldsymbol{\pi}^t$
6: $r_{\text{actual}}^t \leftarrow \Gamma^t / \Gamma_{\max}$ (norm. throughput)
7: **for** $k \in \mathcal{A}$ **do**
8: **if** $k == a^t$ **then**
9: $\hat{r}_k^t \leftarrow r_{\text{actual}}^t$
10: **else**
11: Estimate airtime ($\hat{\tau}_k$) & rate ($\hat{\nu}_k$)
12: $\hat{r}_k^t \leftarrow \hat{\tau}_k \cdot \hat{\nu}_k$
13: **end if**
14: **end for**
15: **// Update internal regret matrix**
16: **for** $k \in \mathcal{A}$ **do**
17: $Q_{a^t \to k} \leftarrow \max\left(0, \lambda \cdot Q_{a^t \to k} + (\hat{r}_k^t - \hat{r}_{a^t}^t)\right)$
18: **end for**
19: **// Update preference vector**
20: **for** $k \in \mathcal{A}$ **do**
21: **if** $k \neq a^t$ **then**
22: $\pi_k^{t+1} \leftarrow \frac{1}{\mu}[Q_{a^t \to k}]_+$
23: **else**
24: $\pi_k^{t+1} \leftarrow 1 - \frac{1}{\mu} \sum_{k \neq a^t} [Q_{a^t \to k}]_+$
25: **end if**
26: **end for**
27: **end for**

C. Solution proposal

Our proposed solution, detailed in Alg. 1, is based on the independent regret-matching algorithm of [5]. The core idea of regret-matching is that each agent iteratively maintains a cumulative swap-regret matrix $\boldsymbol{Q}$, where each entry $Q_{j \to k}$ includes the estimated cumulative gain of having played action k instead of action j. In every iteration t, agent n selects the action that maximizes a preference vector $\boldsymbol{\pi}^t$, i.e., $a_n^t = \arg\max_{k \in \mathcal{A}}, \boldsymbol{\pi}^t$ (Alg. 1, line 5). Unlike typical regret-matching, which adopts a mixed strategy based on a probability distribution, here we use a pure strategy to mitigate the negative effects of random exploration and provide better network stability.

At the end of an iteration, each agent receives a reward r_n^t for the played action (Alg. 1, line 6). Crucially, to enable swap-regret minimization, performance estimates (Alg. 1, lines 11-12) are used to update the matrix $\boldsymbol{Q}$ (Alg. 1, line 17), which effectively drives the update of the preference vector $\boldsymbol{\pi}^{t+1}$. Notice that those preferences are updated in proportion to the internal regrets stored in $\boldsymbol{Q}$, allowing for moving away from suboptimal actions and choosing those that offer higher rewards. In addition, a decay factor λ is applied to better handle non-stationary and avoid relying on too past information. The normalization of the preferences is done using a parameter μ (originally used as *inertia* in [5]), which ensures that all the values sum to 1 and that no preference value becomes negative (Alg. 1, lines 20-26). For that, μ is set to $2(|\mathcal{A}| - 1)$ to ensure that it is strictly greater than the maximum possible sum of positive regrets that an agent could ever accumulate in a single step.

Estimated reward calculation: We define the actual reward of the played action a as the normalized throughput it has experienced, i.e., $r^t_{\text{actual}} = \Gamma^t/\Gamma_{\max}$. To estimate the hypothetical reward of the remaining unplayed actions in iteration ($\forall k \neq a^t \in \mathcal{A}$), we propose an estimator that relies on the "good faith" of the other agents, thus aiming at achieving CE. In particular, we estimate the reward of action n based on its potential performance:

$$\hat{r}^t_k = \hat{\tau}^t_k \cdot \hat{\nu}^t_k, \tag{3}$$

where $\hat{\tau}^t_k$ estimates the airtime when using sensitivity S_k (*can we ignore the others?*) and power P_k (*are the others ignoring us?*) and $\hat{\nu}(k)$ is the estimated bit rate for action k. The latter is extracted from 802.11 tables according to the best Modulation and Coding Scheme (MCS) that supports the estimated Received Signal Strength Indicator (RSSI) at the station ($\hat{\text{RSSI}}_{\text{sta}}$).

The estimated airtime captures the non-linear interactions of CSMA/CA, specifically addressing contention, starvation (fairness), and the capture effect. It is computed as

$$\hat{\tau}_k = \frac{\eta}{\psi_{\text{cont}} \cdot \psi_{\text{fair}}}, \tag{4}$$

where $\psi_{\text{cont}} = 1 + \sum_{m \in \mathcal{N}} \mathbb{I}(\hat{\text{RSSI}}_m \geq S_k)$ is a contention term that represents the number of devices sharing the medium (which depends on the estimated RSSI from node $m \neq n$). ψ_{fair} is a fairness penalty used to prevent hidden nodes. It is set to ω if starvation is detected due to asymmetric interactions, i.e., when agent n cannot detect neighbor m ($\hat{\text{RSSI}}_m < S_k$) but in turn transmits with enough power to silence it ($\hat{\text{RSSI}}_n \geq$ CCA), the agent is causing starvation. Otherwise, it is set to 1. Finally, $\eta = \mathbb{I}(\hat{\gamma}_k > \text{CE})$ ensures that the estimated Signal-to-Interference-plus-Noise Ratio (SINR) based on the power used by the others does not fall below the capture effect threshold (which would potentially lead to packet losses).

The behavior of the considered algorithm is tightly coupled with the accuracy of the reward estimator, provided that the cumulative regret matrix $\boldsymbol{Q}$ is based on the difference between the real revealed reward r_{actual} and the hypothetical estimates $\hat{r}_k$. Specifically, underestimating the potential of unplayed actions may still trap the system in a suboptimal equilibrium. Conversely, overestimating rewards encourages exploration, potentially inducing instability and unfairness. Nevertheless, the estimator does not need to be perfect, as its primary goal is to serve as a heuristic for agents, filling the gap between complete decentralization and centralization.

IV. Performance Evaluation

The evaluation of the proposed algorithm is performed using Komondor [19],[1] a Wi-Fi simulator that includes Machine Learning (ML) agents for driving the optimization of various features, including SR. We consider a scenario comprising two BSSs, each including an AP and a STA (see Fig. 2). This scenario allows for a more precise study of game-theoretic phenomena, given that the interactions between the two players can be easily analyzed. Furthermore, this fits 11bn's MAPC, which considers coordinated transmissions from two APs only. More BSSs will be considered in our future work.

[1]For the sake of openness and reproducibility, all the source code used in this paper is open and can be accessed at https://github.com/wn-upf/Komondor/tree/internal_regret_minimization (commit: 0fb1be3).

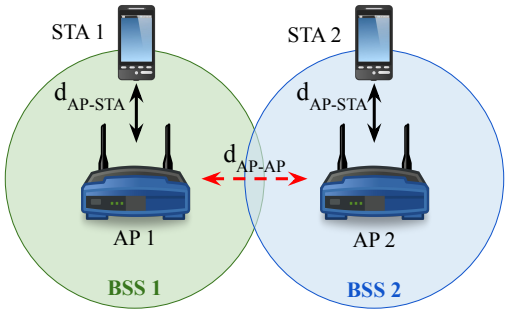


Figure 2: Considered 2-BSS scenario.

Table I: Simulation parameters.

Parameter	Description	Value
T_{sim}	Simulation time	100 s
S	Number of simulated random deployments	100
F_c	Carrier frequency	5 GHz
GI	Guard Interval	3.2 μs
B	Transmission bandwidth	20 MHz
MCS	MCS indices	0-11
$\mathcal{P}^{\text{Noise}}$	Noise power	-95 dBm
$\mathcal{P}_{tx,\max}$	Default transmit power	20 dBm
CCA	Default CCA threshold	-82 dBm
N_{ss}	Single-user spatial streams	1
$G^{\text{TX/RX}}$	Transmitter/receiver antenna gain	0/0 dBi
CE	Capture effect threshold	10 dB
PL	Path loss model and parameters	See [17]
$\text{TXOP}_{\max}$	TXOP duration limit	5.484 ms
$\text{A-MPDU}_{\max}$	A-MPDU size	64
L_D	Length of data packets	1500 bytes
$\mathcal{T}$	Traffic model	Full-buffer
$\mathbb{T}$	Traffic type	Downlink (DL)
CW_0	Initial Contention Window (CW)	16
$\text{CWE}_{\text{min / max}}$	Min./Max. CW exponent	1/5
$\mathcal{S}$	CST values	$\{-62, -72, -82\}$ dBm
ζ	Transmit power values	$\{5, 10, 15, 20\}$ dBm
Δ	Iteration duration	0.5 s
ε_0	Initial exploration coefficient	0.1
$\varepsilon(t)$	Exploration coefficient	$\varepsilon_0/\sqrt{t}$
ω	Fairness penalty	$4(2 \cdot N)$
λ	Discount factor	0.95

Specifically, we compare three different approaches that both agents can take:

- Baseline (`static`): Standard 802.11 operation with fixed parameters ($P = 20$ dBm, $S = -82$ dBm).
- External Regret (ε`-greedy`): A standard MAB approach (ε-greedy) that seeks to maximize individual reward, converging to NE. In ε-greedy, each agent selects a random action with probability ε, while the rest of the times $(1 - \varepsilon)$ exploits the action that has shown the highest payoff so far.
- Internal Regret (`regret-matching`): The proposed regret-matching algorithm described in Section III, converging to CE.

The main simulation parameters are included in Table I.

A. Toy Scenarios

To understand the behavior of agents using external or internal regret minimization strategies, we start with concrete configurations of the deployment depicted in Fig. 2, which allows the definition of two different learning problems:[2] *i*) *strong equilibrium* (for $d_{\text{AP-AP}} = 5$ m and $d_{\text{AP-STA}} = 2$ m) and *ii*) *weak equilibrium* (for $d_{\text{AP-AP}} = 4$ m and $d_{\text{AP-STA}} = 2$ m). Varying the positions of the devices leads to different spatial interactions among BSSs. In the first case (strong equilibrium), both APs can successfully transmit in parallel if they increase their sensitivity (e.g., using $S = -72$ dBm and $P = 20$ dBm). Therefore, the strategy maximizing the overall performance is the same as for maximizing the individual performance. As a result, finding the optimal solution is straightforward for greedy agents. This does not occur in the weak-equilibrium case, where simultaneous transmissions at the maximum transmit power would lead to collisions at both STAs. In that case, the optimal action consists of lowering the transmit power to minimize the interference generated towards the other BSS ($P = 10$ dBm) and increasing the CST to avoid contention ($S = -72$ dBm).

The mean average throughput achieved across the two BSSs during the simulations of the two cases (strong vs. weak equilibrium) is shown in Fig. 3 for each considered approach. As shown, in the first case (strong equilibrium), the static (default) configuration leads to an average throughput of ∼60 Mbps because the two BSSs alternate the access to the medium. In that case, both the external (ε-greedy) and internal (regret-matching) regret minimization approaches allow reusing the space and performing simultaneous transmissions. It is worth noting that, because it sticks to the best action and performs to further exploration, regret-matching leads to slightly higher performance than ε-greedy, which in turn continues to explore random actions even after finding the best action set. When it comes to the weak equilibrium case, we find that the ε-greedy approach gets stuck in the default performance, provided that the agents play aggressively and are not capable of finding the best configuration. In turn, when using regret-matching, the agents are able to discover the best-performing action, thus leading to a mean average throughput above 80 Mbps.

To better understand the behavior of each learning algorithm, Fig. 4 shows the actions selected by each agent in every iteration. Starting with the strong equilibrium scenario, ε-greedy (Fig. 4a) discovers the best action (A_2), but continues exploring suboptimal actions throughout the simulation. Regret-matching (Fig. 4b), in contrast, quickly converges to the best action and does not perform any other exploration. In this case, regret-matching spends several iterations stuck in a suboptimal arm (A_4), but soon switches to the true optimal arm as a result of having accumulated positive regret for such an action during several consecutive iterations. When it comes to the weak equilibrium case, ε-greedy gets stuck in a suboptimal arm (see Fig. 4c), whereas regret-matching quickly converges to the optimal one (see Fig. 4d).

[2]For the sake of analysis, in this setup we limit the actions to two different values of sensitivity (-72 dBm, -82 dBm) and power (10 dBm, 20 dBm).

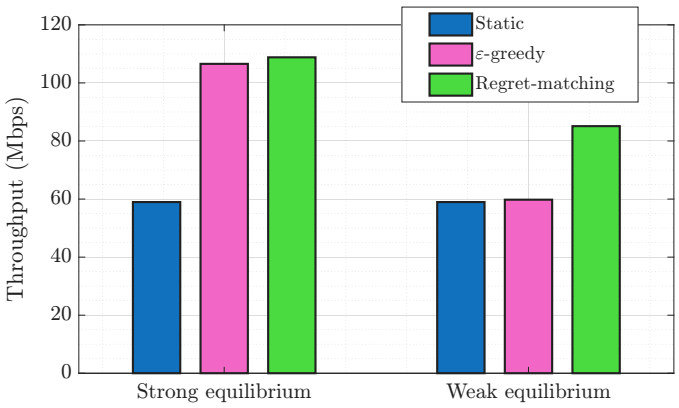
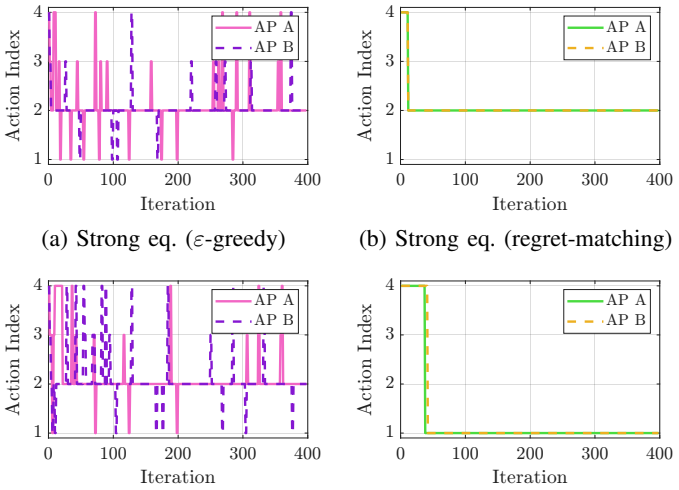


Figure 3: Mean average throughput (in Mbps) obtained by the two BSSs in the scenario encompassing strong and weak equilibrium situations.



(a) Strong eq. (ε-greedy) (b) Strong eq. (regret-matching)

(c) Weak eq. (ε-greedy) (d) Weak eq. (regret-matching)

Figure 4: Actions selected by the two BSSs per iteration in each scenario. Actions available (CST, transmit power): A_1 = {-72, 10} dBm, A_2 = {-72, 20} dBm, A_3 = {-82, 10} dBm, A_4 = {-82, 20} dBm.

B. Random Deployments

To showcase the potential of the regret-matching solution, we now consider random deployments comprising the two studied BSSs from Fig. 2. In particular, we study the effect of inter-AP interference, accounting for different distances between BSSs based ($d_{\text{AP-AP}}$). The position of the STAs is also randomly selected around their APs, with $d_{\text{AP-STA}} = [3-5]$ m. The throughput results (mean and minimum across BSSs) are depicted in Fig. 5.

As shown, regret-matching consistently improves the performance of the two BSSs across different deployment setups. For shorter distances, increasing SR is challenging, hence the gains are limited. However, as the distance between BSSs increases, hence the inter-AP interference decreases, the SR performance gains are more significant. When it comes to the minimum throughput, which embodies fairness in all BSSs,

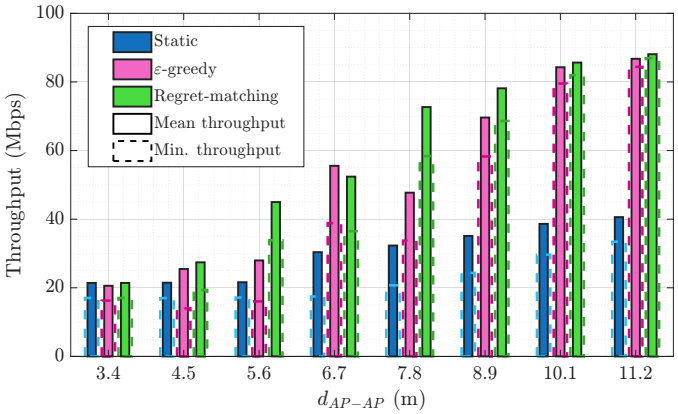


Figure 5: Mean (solid bars) and minimum (dashed bars) average throughput achieved across random deployments.

regret-matching offers better performance in most cases, thus underscoring its game-theoretic design.

V. Conclusions

As Wi-Fi networks evolve towards the upcoming IEEE 802.11bn (Wi-Fi 8), the management of inter-BSS interference has become an important challenge for reliability. While the standard is shifting toward centralization with MAPC (i.e., Co-SR), such architectures impose significant signaling overheads that increase complexity and limit scalability. In this paper, we proposed a learning-based method that leverages Game Theory to bridge the gap between decentralized decision-making and global optimality. In particular, our method adopts an internal regret minimization approach (based on regret-matching) to achieve CE. With this, we aimed to achieve implicit coordination among BSSs, but without having to exchange a single bit for coordination purposes. Our simulation results confirm that this approach has practical potential, as it has been shown to overcome the limitations of state-of-the-art external regret minimization methods. As the IEEE 802.11 groups define the specifications for next-generation Wi-Fi networks, algorithms grounded in internal regret minimization offer a compelling alternative to scalable, high-efficiency, and flexible optimization solutions, unlocking the full potential of the unlicensed spectrum.

Acknowledgments

This paper is supported by the CHIST-ERA Wireless AI 2022 call MLDR project (ANR-23-CHR4-0005), partially funded by AEI under project PCI2023-145958-2, by TRUE-Wi-Fi PID2024-155470NB-I00 and Wi-XR PID2021-123995NB-I00 (MCIU/AEI/FEDER,UE), by MCIN/AEI under the Maria de Maeztu Units of Excellence Programme (CEX2021-001195-M), and AGAUR ICREA Academia 00077.

References

[1] G. Geraci, F. Meneghello, F. Wilhelmi, D. Lopez-Perez, I. Val, L. G. Giordano, C. Cordeiro, M. Ghosh, E. Knightly, and B. Bellalta, "Wi-Fi: Twenty-five years and counting," *Proceedings of the IEEE*, 2026.

[2] I. Jamil, L. Cariou, and J.-F. Hélard, "Novel learning-based spatial reuse optimization in dense WLAN deployments," *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 1, p. 184.

[3] F. Wilhelmi, C. Cano, G. Neu, B. Bellalta, A. Jonsson, and S. Barrachina-Muñoz, "Collaborative spatial reuse in wireless networks via selfish multi-armed bandits," *Ad Hoc Networks*, vol. 88, pp. 129–141, 2019.

[4] "IEEE standard Draft 1.1 for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. Amendment 6: Enhancements for ultra high reliablity (UHR)," *IEEE P802.11bn/D1.1*, 2025.

[5] S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, vol. 68, no. 5, pp. 1127–1150, 2000.

[6] F. Wilhelmi, S. Barrachina-Muñoz, C. Cano, I. Selinis, and B. Bellalta, "Spatial reuse in IEEE 802.11 ax WLANs," *Computer Communications*, vol. 170, pp. 65–83, 2021.

[7] E. de Carvalho Rodrigues, A. Garcia-Rodriguez, L. G. Giordano, and G. Geraci, "On the latency of IEEE 802.11 ax WLANs with parameterized spatial reuse," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–6.

[8] F. Wilhelmi, B. Bellalta, S. Szott, K. Kosek-Szott, and S. Barrachina-Muñoz, "Coordinated Multi-Armed Bandits for Improved Spatial Reuse in Wi-Fi," in *2025 IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN)*. IEEE, 2025, pp. 1–6.

[9] F. Wilhelmi, B. Bellalta, C. Cano, and A. Jonsson, "Implications of decentralized Q-learning resource allocation in wireless networks," in *2017 ieee 28th annual international symposium on personal, indoor, and mobile radio communications (pimrc)*. IEEE, 2017, pp. 1–5.

[10] P. E. Iturria-Rivera, M. Chenier, B. Herscovici, B. Kantarci, and M. Erol-Kantarci, "Cooperate or not cooperate: Transfer learning with multi-armed bandit for spatial reuse in Wi-Fi," *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 2, pp. 351–369, 2024.

[11] B. Yin, K. Yamamoto, T. Nishio, M. Morikura, and H. Abeysekera, "Learning-based spatial reuse for WLANs with early identification of interfering transmitters," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 151–164, 2019.

[12] Y. Huang and K.-W. Chin, "A deep Q-network approach to optimize spatial reuse in WiFi networks," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 6, pp. 6636–6646, 2022.

[13] Y. Song, C. Zhang, and Y. Fang, "Joint channel and power allocation in wireless mesh networks: A game theoretical perspective," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 7, pp. 1149–1159, 2008.

[14] M. Bloem, T. Alpcan, and T. Başar, "A stackelberg game for power control and channel allocation in cognitive radio networks," in *Proceedings of the 2nd international conference on Performance evaluation methodologies and tools*, 2007, pp. 1–9.

[15] C. K. Tan, M. L. Sim, and T. C. Chuah, "Game theoretic approach for channel assignment and power control with no-internal-regret learning in wireless ad hoc networks," *IET communications*, vol. 2, no. 9, pp. 1159–1169, 2008.

[16] S. Maghsudi and S. Stańczak, "Channel selection for network-assisted d2d communication via no-regret bandit learning with calibrated forecasting," *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1309–1322, 2014.

[17] F. Wilhelmi, S. Barrachina-Munoz, B. Bellalta, C. Cano, A. Jonsson, and G. Neu, "Potential and pitfalls of multi-armed bandits for decentralized spatial reuse in WLANs," *Journal of Network and Computer Applications*, vol. 127, pp. 26–42, 2019.

[18] A. Blum and Y. Mansour, "From external to internal regret." *Journal of Machine Learning Research*, vol. 8, no. 6, 2007.

[19] S. Barrachina-Munoz, F. Wilhelmi, I. Selinis, and B. Bellalta, "Komondor: A wireless network simulator for next-generation high-density WLANs," in *2019 Wireless Days (WD)*. IEEE, 2019, pp. 1–8.