Federated Non-Stochastic Multi-Armed Bandit for Channel Sensing in Cognitive Radio Systems

Kinda Khawam, Farah Yassine, Samer Lahoud, Yujie Tang, Dominique Quadri, Steven Martin

Federated Non-Stochastic Multi-Armed Bandit for Channel Sensing in Cognitive Radio Systems

Kinda Khawam[1,2], Farah Yassine[1], Samer Lahoud[3], Yujie Tang[3], Dominique Quadri[1], and Steven Martin[1]
[1] ROCS, LISN, Université Paris Saclay, 91190 Gif-sur-Yvette, France
[2] Université de Versailles Saint-Quentin-en-Yvelines, 78000 Versailles, France
[3] Faculty of Computer Science, Dalhousie University, Halifax, NS, B3H 4R2 Canada

***Abstract*—This work proposes a novel approach to channel sensing in cognitive radio systems, drawing inspiration from reinforcement learning theory. While the adversarial Multi-Armed Bandit framework is commonly used to manage diverse channel sampling, it struggles to accurately assess resource occupancy due to geographical variations in channel availability across devices. To address this limitation, collaboration among devices is essential and can be effectively achieved through Federated Learning (FL). FL integrated into a Multi-Armed Bandit framework addresses key challenges, including data heterogeneity, the high cost of data centralization, privacy concerns, and biased learning. We enhance the widely-used Exponential-weight algorithm for exploration and exploitation (EXP3) by incorporating federation, allowing learning devices to collectively identify channels with less interference. Our simulation results demonstrate the effectiveness of the federated EXP3 (F-EXP3) algorithm by comparing it with the traditional EXP3 and the federated Upper Confidence Bound (UCB). The experiments reveal that F-EXP3 overcomes the limitations of individual learning, leading to superior channel selection performance.**

***Index Terms*—Federated Learning, Non-Stochastic Multi-Armed Bandit, Cognitive Radio, Channel Sensing.**

I. Introduction

Federated Learning (FL) enables decentralized Machine Learning, enhancing privacy and cost-effectiveness [1], [2], particularly when data is spread across various devices and centralizing it to reduce learning error would be too expensive [3]. An innovative extension of FL is the Federated Multi-Armed Bandit (FMAB), which combines FL with sequential decision-making [4]. Unlike traditional Multi-Armed Bandit (MAB) models focused solely on the exploration-exploitation trade-off, the federated bandit approach also addresses challenges like data heterogeneity, the high cost of data centralization, and the need for privacy protection.

In FL, data available to individual clients typically comes from non-IID (independent and identically distributed) distributions, requiring collaboration among clients to derive meaningful insights from the aggregated global model. As in traditional federated settings, a central server orchestrates the learning process among clients. To address privacy concerns, clients do not transmit raw data; instead, they only send the learned model weights to the central server for aggregation.

In this paper, we examine a cognitive radio framework where a base station (BS) is responsible for selecting the least occupied channel from a designated set within its coverage area. Channel availability is known to vary across different geographical locations, with the overall availability often treated as the ground truth—an average across the entire coverage area. However, the BS, being fixed at a specific location, cannot independently determine this global availability. To overcome this limitation, a common strategy is to delegate the task to randomly positioned devices, such as mobile phones or IoT devices, dispersed throughout the coverage area to assess channel availability. The data collected by these devices is then aggregated at the BS, offering a more comprehensive view of channel usage.

The existing literature on Federated Bandit algorithms predominantly addresses the stochastic bandit setting, as demonstrated in foundational works such as [4], [5], and [6]. Some extensions explore cases where the expected reward is a linear function of the selected arm, as seen in [7] and [8], or scenarios involving an infinite number of arms, as discussed in [9]. However, these methods often fall short when dealing with realistic scenarios where clients' datasets are non-IID, leading to diverse local objectives. In such cases, particularly in non-stochastic environments, the adversarial Multi-Armed Bandit (MAB) framework is more appropriate [10]. Unlike stochastic MAB, the adversarial MAB framework does not depend on any predefined distribution of rewards, making it better suited for handling variability in clients' data.

In this paper, we propose F-EXP3, a federated version of the EXP3 algorithm designed for non-stochastic settings. Our results show that F-EXP3 performs comparably to centralized EXP3, with bounded regret relative to the best fixed choice. It also effectively identifies optimal channels, mitigating biases from devices geographical locations. Additionally, we demonstrate F-EXP3's advantages over the federated stochastic MAB approach, particularly the federated Upper Confidence Bound (UCB) [4].

The remainder of the paper is organized as follows: section II introduces the network model and formulates the problem. Section III describes the F-EXP3 algorithm. Section IV presents the simulation results. Finally, section V outlines the conclusion.

II. Network Model and Problem Formulation

A BS in a cognitive radio network aims to autonomously select the most suitable channels with the minimum received interference power level from a predefined set, while remaining responsive to variations in shared resource utilization

and radio channel quality. Given that the BS is stationary and channel availability fluctuates across different locations, collaborating with randomly positioned devices in a federated manner can significantly enhance learning efficiency.

In the Federated Multi-Armed Bandit (FMAB) framework depicted in Fig. 1, the central server, referred to as the central base station (BS^*), aims to identify the least utilized channel among J neighboring BSs. This is achieved through interactions with clients, which are individual devices. Each device, denoted as m where $m \in 1, ..., M$, operates on K channels, analogous to the arms in a standard Multi-Armed Bandit (MAB) setup. All clients share the same set of K arms, known as local arms at this level [4]. Each client interacts exclusively with its own local MAB model, and does not communicate directly with other clients. Although the learning objective is centered on the server's MAB model, known as the global model, the server does not have direct access to it; instead, it relies on feedback from clients about their local observations.

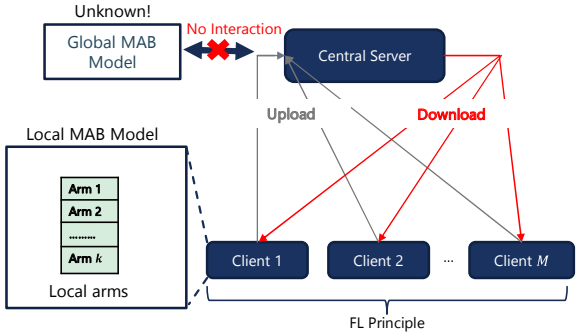

Fig. 1. System model for federated MAB.

Within a time horizon T, each client m aims to select the most available channel k (where $1 \leq k \leq K$) to maximize its expected cumulative reward G^m shown below:

$$G^m := \sum_{t=1}^{T} r^m_{k^m_t}(t), \tag{1}$$

where the channel k_t is randomly chosen according to a distribution governed by EXP3, as described in Algorithm 1.

We denote the reward as $r^m_k \in (0, 1]$, for sensing channel k by device m as follows:

$$r^m_k = 1 - \frac{P^m_k}{P^{m^{max}}_k}. \tag{2}$$

P^m_k represents the received power from other network(s) $j \in 1, .., J$, on channel k sensed by device m, expressed as:

$$P^m_k = P_0 \cdot \sum_{j \in J} G_j \cdot x_{k,j} \cdot y_{k,j} \cdot (\frac{1}{d^k_{m,j}})^{\beta} + P_N. \tag{3}$$

P_0 denotes the transmitted power per channel, while P_N represents the thermal noise power per channel. Additionally, $y_{k,j}$ are i.i.d. random variables (with unit mean) that depict the impact of fast fading experienced on channel k, as sensed by BS j. These variables follow an exponential distribution to model Rayleigh fading [11]. G_j is the antenna gain of BS j, β represents the path-loss factor, and $d^k_{m,j}$ is the distance between BS j and device m sensing channel k. Finally, the variable $x_{k,j}$ serves as a binary indicator defined as follows:

$$x_{k,j} = \mathbb{1}_{\{BS\, j \ selects\, channel\, k\}}. \tag{4}$$

$P^{m^{max}}_k$ represents the highest value of P^m_k that can be achieved expressed as $P^{m^{max}}_k = P_0 \cdot \sum_{j \in J} G_j + P_N$. It serves as a normalization factor to ensure that the reward remains below 1.

The cumulative reward is usually compared with the largest reward given by the best fixed choice denoted by g^m and expressed below:

$$g^m := \max_{k=1...K} \sum_{t=1}^{T} r^m_k(t). \tag{5}$$

The regret R^m_T is defined as the difference between g^m and G^m of any device m as follows:

$$R^m_T := g^m - \mathbb{E}[G^m], \tag{6}$$

where the averaging is taken w.r.t. the probabilistic choices of device m.

III. Federating the EXP3 MAB

We begin by introducing the standard Exponential-weight Algorithm for Exploration and Exploitation (EXP3) [12], which is utilized by each device m in the network. In this setup, each device m, at each time step t, selects a channel k from a set of K available channels, and receives a corresponding reward $r^m_k(t)$. We consider a vector space $\{0, 1\}^K$, where the strategy space for each device m is defined as $S_m \in \{0, 1\}^K$ of size K. Following this, we present our Federated EXP3 (F-EXP3) algorithm, which operates at the central BS^* that integrates learning feedback from all participating devices.

A. The EXP3 Algorithm

The EXP3 algorithm, detailed in Algorithm 1, was originally introduced in [13] to tackle sequential allocation problems. At each time step t within a time horizon T, a device m executes the EXP3 algorithm. Initially, each device m assigns an equal weight to each channel, setting $\omega_k(1) = 1$ for all channels $k \in 1, .., K$. Subsequently, the device selects a channel k^m_t at random, using a probability distribution calculated as described in line 2 of Algorithm 1. This probability distribution is a combination of a uniform distribution, which ensures exploration of all K channels, and an exponential distribution that assigns probabilities proportional to the estimated cumulative reward of each channel.

B. The F-EXP3 Algorithm

The F-EXP3 algorithm, outlined in Algorithm 2, is executed in parallel by the devices and the central server. At each iteration t, each device m runs the EXP3 algorithm (Algorithm 1) to select a channel, and update its weights $w^m(t)$ (lines 4 to 6). These updated weights are sent to the central server BS^*, which aggregates them by summing (lines 8 and 9). The aggregated weights are then sent back to the devices as the global initial weights for the next iteration $t+1$ (line 10).

Notably, only the weights are exchanged between the devices and the central server in F-EXP3, enhancing privacy compared to previous methods like in [4], where actual rewards are transmitted.

Algorithm 1 EXP3 applied by any device m

1: **INPUT:** γ a real parameter in $(0, 1]$, and $\omega(t)$
2: **for** i = 1 to K **do**

$$p_i^m(t) = (1-\gamma)\frac{\omega_i(t)}{\sum_{k=1}^{K}\omega_k(t)} + \frac{\gamma}{K} \tag{7}$$

3: **end for**
4: Device m selects randomly channel i_t^m based on the probabilities $p_1^m(t), .., p_K^m(t)$
5: Device m receives reward $r_{i_t^m}^m(t)$ for drawn channel i_t^m
6: **for** i = 1 to K **do**

$$\hat{r}_i^m(t) = \begin{cases} \frac{r_i^m(t)}{p_i^m(t)} & \text{if } i = i_t^m \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

$$\omega_i^m(t+1) = \omega_i(t)\exp(\frac{\gamma \hat{r}_i^m(t)}{K}) \tag{9}$$

7: **end for**
8: **OUTPUT:** $\omega^m(t+1)$

Algorithm 2 F-EXP3 applied by the central server BS^*

1: **INPUT:** Let γ be a real parameter in $(0, 1]$
2: **INITIALIZATION** Set initial channel weights $\omega_i(1) = 1$ for every channel $i \in \{1, .., K\}$
3: **for** t = 1 to T **do**
4: **for** m=1 to M **do**
5: $\omega^m(t)$ = EXP3(γ, $\omega(t)$), using Algorithm 1
6: **end for**
7: BS^* receives weights $\omega^m(t)$ from devices $m = 1, \ldots, M$
8: **for** i=1 to K **do**

$$\hat{\omega}_i(t) = \frac{1}{M}\sum_{m=1}^{M}\omega_i^m(t) \tag{10}$$

9: **end for**
10: $\omega(t+1) = \hat{\omega}(t)$
11: **end for**

Additionally, F-EXP3 is fully decentralized, computationally efficient, and guarantees bounded regret, ensuring strong performance as detailed in Theorem3.1.

Theorem 3.1: Regardless of the reward functions, F-EXP3 guarantees that the expected regret is bounded by:

$$R_T \leq (e-1)\gamma g + \frac{K\log(K)}{\gamma}. \tag{11}$$

R_T is the mean regret across all devices, defined as $R_T = \frac{1}{M}\sum_{m=1}^{M} R_T^m$, while g is a bound on the maximum cumulative reward, limited by T. The proof is provided in the Appendix.

IV. Simulation Results

In this section, we demonstrate the effectiveness of the proposed F-EXP3 algorithm compared to the EXP3 algorithm and the Federated Stochastic UCB from [4], where the local models are stochastic realizations of the global model. We evaluate performance based on the probability of selecting the best channel and the regret across various scenarios.

We consider a cognitive radio network with $M = 100$ devices uniformly distributed around a central BS^*. The network has K available channels, and the central BS^* collaborates with its devices to select the optimal channel. The devices also monitor the channel utilization of J neighboring networks, with each neighboring base station BS_j located 1 km apart. Each BS_j assigns channels based on a uniform distribution, with a usage ratio of $\hat{x}_{k,j} = \mathbb{E}(x_{k,j})$, which is unknown to the learning devices. While neighboring cells use resources randomly, BS^* employs intelligent learning.

The simulation parameters are as follows: the mean antenna gain G_j for each neighboring base station (BS_j) and the transmission power (P_0) are both normalized to 1, with the noise power set to $P_N = P_0 \times 10^{-4}$. The path-loss exponent is $\beta = 3$. We use the optimal value of γ, as derived in [12], given by $\gamma = \sqrt{\frac{K\log(K)}{(e-1)\cdot T}}$. The simulation results are generated using a discrete event simulator implemented in SimPy [14].

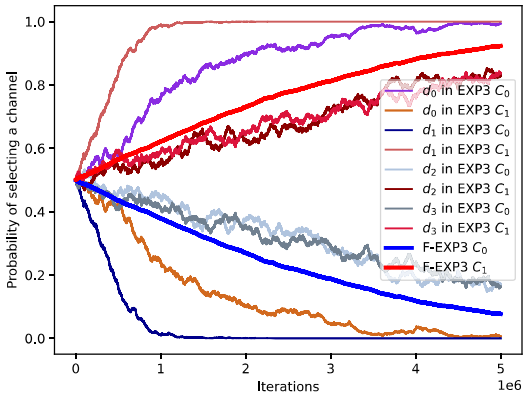

Fig. 2. F-EXP3 vs. EXP3: variation in the probability distribution of selected channels with utilization rates $\hat{x}_{0,0} = 40\%$, $\hat{x}_{1,0} = 60\%$, $\hat{x}_{0,1} = 80\%$, and $\hat{x}_{1,1} = 20\%$.

Figs. 2 and 3 show the variation in the probability distribution, as defined in (7), over $T = 5000000$ iterations. We compare the performance of F-EXP3 with EXP3 for two neighboring BSs ($J = 2$, BS_0 and BS_1) across four devices:

d_0, d_1, d_2, and d_3. Devices d_0 and d_1 are positioned close to BS_0 and BS_1, respectively, while d_2 and d_3 are randomly placed.

In Fig. 2, BS_0 and BS_1 use the $K = 2$ available channels with specific proportions: BS_0 uses channel 0 (C_0) with $\hat{x}_{0,0} = 40\%$ and channel 1 (C_1) with $\hat{x}_{1,0} = 60\%$, while BS_1 uses C_0 with $\hat{x}_{0,1} = 80\%$ and C_1 with $\hat{x}_{1,1} = 20\%$. We notice that Only EXP3 applied to device d_1, which is close to BS_1, outperforms F-EXP3. Device d_0, despite being near BS_0, fails to identify the best channel due to the fair sharing of both channels used by its neighboring BS_0, and the far distance from BS_1. For devices d_2 and d_3, EXP3 shows slower convergence compared to F-EXP3.

In Fig. 3, we adjust the channel usage ratios of neighboring BSs to $\hat{x}_{0,0} = 50\%$ and $\hat{x}_{0,1} = 80\%$. All devices (d_0, d_1, d_2, d_3), and F-EXP3 successfully identify the optimal channel C_1. We observe F-EXP3 outperforming EXP3 applied to d_0, d_2, and d_3, while EXP3 applied to d_1 converges faster due to its proximity to BS_1, which heavily utilizes C_0.

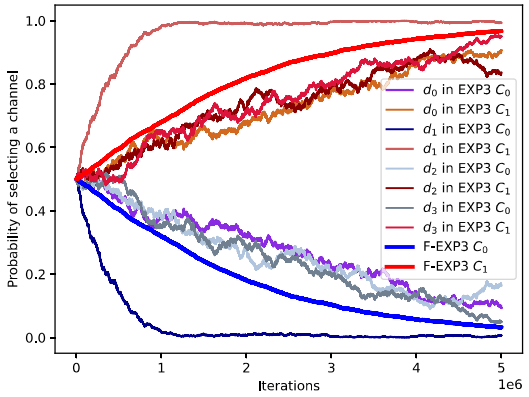

Fig. 3. F-EXP3 vs. EXP3: variation in the probability distribution of selected channels with utilization rates $\hat{x}_{0,0} = 50\%$, $\hat{x}_{1,0} = 50\%$, $\hat{x}_{0,1} = 80\%$, and $\hat{x}_{1,1} = 20\%$.

We conclude that relying on a single device to execute the EXP3 algorithm is not a reliable solution.

In Fig. 4, we compare the complexity of F-EXP3 to EXP3 by examining the number of iterations required to achieve a probability of 0.8 for selecting the optimal channel C_1, using the same channel utilization ratios as in Fig.3. The results show that EXP3 applied to d_1 requires the fewest iterations ($\leq 0.5 \cdot 10^6$) due to its proximity to BS_1, which predominantly uses C_0. However, F-EXP3 requires significantly fewer iterations ($\leq 1.5 \cdot 10^6$) than EXP3 applied to d_0, d_2, and d_3. Additionally, EXP3 shows a high error interval for d_0, d_2, and d_3 due to fluctuating channel conditions (channel fading), highlighting the benefit of device cooperation in F-EXP3 for improving channel selection accuracy in non-stochastic wireless environments.

In Fig. 5, we examine the cumulative regret per device, as calculated in (6), for both EXP3 and F-EXP3 over the time horizon T, using the same channel utilization ratios as in Fig. 3. F-EXP3 consistently results in lower cumulative regret for devices d_0, d_2, and d_3 compared to EXP3, with regret staying below the theoretical bound. However, d_1 shows higher cumulative regret with F-EXP3 than with EXP3, due to its proximity to BS_1, which heavily utilizes C_0. In this optimal position, d_1 can effectively identify the best channel using EXP3 alone, as seen in Figs. 2 and 3. On the other hand, devices like d_2 and d_3, which are randomly placed, or d_0, which is near BS_0 with balanced channel usage, struggle to find the optimal channel using EXP3 independently. F-EXP3 enables these devices to collaborate by sharing sensing feedback, thereby improving detection accuracy and reducing average cumulative regret (shown by the yellow curve).

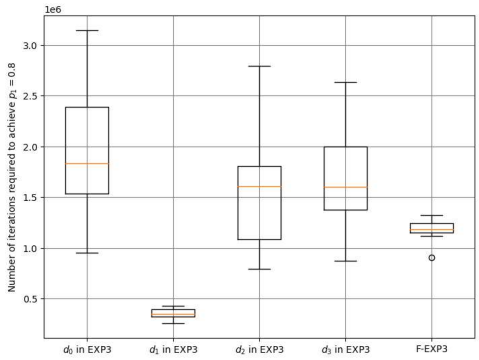

Fig. 4. F-EXP3 vs. EXP3: variation in the number of iterations required to achieve a probability value of 0.8 for selecting C_1, with utilization rates $\hat{x}_{0,0} = 50\%$, $\hat{x}_{1,0} = 50\%$, $\hat{x}_{0,1} = 80\%$, and $\hat{x}_{1,1} = 20\%$.

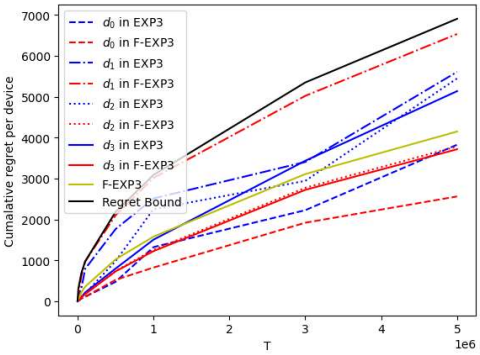

Fig. 5. F-EXP3 vs. EXP3: variation in the cumulative regret with utilization rates $\hat{x}_{0,0} = 50\%$, $\hat{x}_{1,0} = 50\%$, $\hat{x}_{0,1} = 80\%$, and $\hat{x}_{1,1} = 20\%$.

In Fig. 6, we examine how the number of devices involved in the F-EXP3 algorithm influences the iterations needed to achieve a 0.8 probability of selecting the optimal channel (C_1). We analyze four networks with 5, 20, 50, and 100 devices, conducting 20 realizations for each scenario. The results show that as the number of devices increases, fewer iterations are required to reach the target probability. For example, with 100 devices, the median number of iterations is 6.7×10^5, compared to 7.88×10^5 with only 5 devices. Additionally, increasing the number of devices reduces the error bar, highlighting the importance of device distribution.

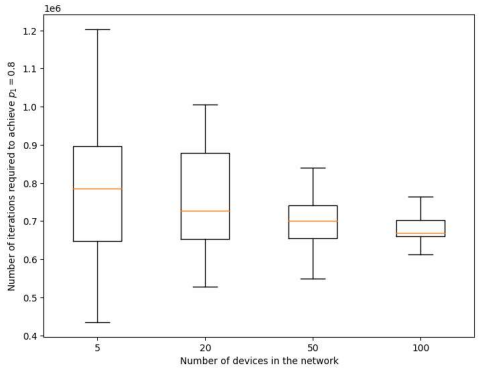

Fig. 6. F-EXP3: variation in the number of iterations required to achieve a probability value of 0.8 for selecting C_1, with utilization rates $\hat{x}_{0,0} = 50\%$, $\hat{x}_{1,0} = 50\%$, $\hat{x}_{0,1} = 80\%$, and $\hat{x}_{1,1} = 20\%$.

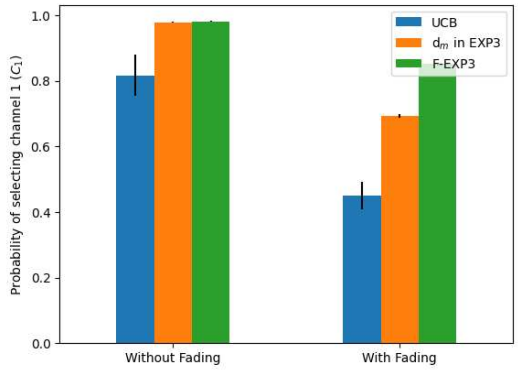

Fig. 7. EXP3 on d_m vs. F-EXP3 vs. Federated UCB: variation in the probability of selecting channel C_1 as a function of the presence of channel fading.

A larger number of uniformly distributed devices leads to a more stable federated learning process, improving convergence to the optimal channel.

In Fig. 7, we compare the performance of the F-EXP3 algorithm with the traditional EXP3, applied to a randomly selected device d_m, and the Federated UCB algorithm [4] in selecting the optimal channel (C_1) over $T = 1,000,000$ iterations across 20 realizations. We assess two scenarios: one without channel fading (stochastic setting) and one with channel fading (non-stochastic setting). The network consists of $J = 4$ base stations (BS_0, BS_1, BS_2, BS_3) and $K = 4$ channels, with each base station selecting channels in the proportions $\hat{x}_{0,j} = 30\%$, $\hat{x}_{1,j} = 10\%$, $\hat{x}_{2,j} = 30\%$, and $\hat{x}_{3,j} = 30\%$. Our results show that F-EXP3 consistently outperforms both EXP3 and Federated UCB in selecting C_1, regardless of channel fading. Notably, Federated UCB is designed for stochastic environments, where it struggles with the unpredictability of fading channels. In contrast, our proposed F-EXP3 algorithm is specifically designed to adapt effectively to the variability of realistic channels. Specifically, in the absence of fading, both F-EXP3 and EXP3 achieve a probability close to 1, whereas Federated UCB stagnates at 0.8. With fading, F-EXP3 maintains a probability of 0.85, surpassing EXP3 at 0.7 and Federated UCB at 0.44. These results highlight the effectiveness of F-EXP3 in adapting to varying channel conditions and improving channel selection performance.

Fig. 8 illustrates the probability of selecting the optimal channel (C_1) over $T = 1,000,000$ iterations across 20 realizations in four different environments: $J = 2$ BSs with $K = 2$ channels, $J = 4$ BSs with $K = 4$ channels, $J = 6$ BSs with $K = 6$ channels, and $J = 8$ BSs with $K = 8$ channels. In each scenario, C_1 is less utilized by each BS_j. As the number of channels increases, UCB struggles to select the less utilized channel (C_1) due to channel fading, with its probability dropping from 0.9 with 2 channels to less than 0.4 with 8 channels. In contrast, F-EXP3 and EXP3 consistently identify C_1 regardless of the number of available channels, with F-EXP3 performing better (e.g., probabilities of 0.83 for F-EXP3 and 0.72 for EXP3 with 4 channels). This demonstrates the effectiveness of F-EXP3 in leveraging device cooperation to select C_1 more efficiently than EXP3. However, as the number of available channels increases, both algorithms see reduced probability values, indicating a need for more iterations with increasing channel number to achieve higher probabilities of selecting the optimal channel.

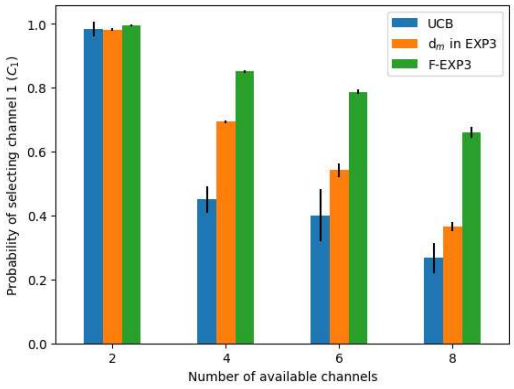

Fig. 8. EXP3 on d_m vs. F-EXP3 vs. Federated UCB: variation in the probability distribution of selecting channel C_1 as a function of the number of available channels.

V. Conclusion

In this work, we introduced a novel framework for the federated multi-armed adversarial bandit problem in channel selection. Our framework addresses key challenges, such as the heterogeneity of local learning objectives due to the non-IID nature of wireless channel sampling, and the lack of access to a global objective for both clients (devices) and the central server (base station). To tackle these issues, we proposed the F-EXP3 algorithm, capable of handling non-stochastic local objectives. Our results demonstrated the superiority of this federated adversarial bandit approach over both centralized and federated stochastic bandit methods. We showed that federating non-stochastic reinforcement learning across diverse devices enhances resource utilization efficiency. Future work will explore applying smart F-EXP3 to all base stations, and

assess its impact on convergence rates. We anticipate that using F-EXP3 across all base stations will increase randomness in reward generation, making the adversarial federated Multi-Armed Bandit even more appropriate.

VI. Appendix

The proof of Theorem 3.1 follows that of the classical proof of EXP3 regret bound in [12]. We denote by $W_t := \frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M}\omega_k^m(t)$. Note that $p_k^m(t)$ in (7) boils down to $p_k(t)$ after weights aggregation. We use the following simple facts:

$$\hat{r}_k^m(t) \leq \frac{1}{p_k^m(t)} = \frac{1}{p_k(t)} \leq \frac{1}{K} \tag{12}$$

$$\sum_{k=1}^{K} \; p_k(t)\cdot \hat{r}_k^m(t) = r_{k_t^m}^m(t) \tag{13}$$

$$\begin{aligned}\sum_{k=1}^{K} p_k(t)\cdot(\hat{r}_k^m(t))^2 &= \sum_{k=1}^{K} r_k^m(t)\cdot\hat{r}_k^m(t)\\ &\leq \hat{r}_{k_t^m}^m(t) = \sum_{k=1}^{K}\hat{r}_k^m(t)\end{aligned} \tag{14}$$

The proof is based on the following sequence of equations:

$$\frac{W_{t+1}}{W_t} = \frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M}\frac{\omega_k^m(t+1)}{W_t} = \frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M}\frac{\omega_k(t+1)}{W_t} \tag{15}$$

$$= \frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M}\frac{p_k^m(t)-\frac{\gamma}{K}}{1-\gamma}\exp\left(\frac{\gamma}{K}(\hat{r}_k^m(t))\right) \tag{16}$$

$$= \frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M}\frac{p_k(t)-\frac{\gamma}{K}}{1-\gamma}\exp\left(\frac{\gamma}{K}(\hat{r}_k^m(t))\right) \tag{17}$$

Equation (16) uses the definition of $p_k^m(t)$ in (7). Then, when $x \leq 1$, we have $e^x \leq 1 + x + (e-2)x^2$; hence, Equation (17) can be rewritten as inequality (18).

$$\begin{aligned}\frac{W_{t+1}}{W_t} \leq \frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M}\frac{p_k(t)-\frac{\gamma}{K}}{1-\gamma}(1+\frac{\gamma}{K}\hat{r}_k^m(t)+\\ (e-2)(\frac{\gamma}{K}\hat{r}_k^m(t))^2)\end{aligned} \tag{18}$$

$$\begin{aligned}&\leq 1 + \frac{\gamma}{M\cdot K(1-\gamma)}\sum_{m=1}^{M} r_{k_t^m}^m(t)\\ &+\frac{(e-2)\gamma^2}{M\cdot K^2(1-\gamma)}\sum_{k=1}^{K}\sum_{m=1}^{M}\hat{r}_k^m(t)\end{aligned} \tag{19}$$

Taking logarithm and since $\log(1+x) \leq x$, (19) implies:

$$\begin{aligned}\log\left(\frac{W_{t+1}}{W_t}\right) \leq &\frac{\gamma}{M\cdot K(1-\gamma)}\sum_{m=1}^{M} r_{k_t^m}^m(t)\\ &+\frac{(e-2)\gamma^2}{M\cdot K^2(1-\gamma)}\sum_{k=1}^{K}\sum_{m=1}^{M}\hat{r}_k^m(t)\end{aligned} \tag{20}$$

Summing from $t=1$ to $t=T$, and owing to (1), we get:

$$\begin{aligned}\log\left(\frac{W_{T+1}}{W_1}\right) \leq &\frac{\gamma}{K(1-\gamma)}\frac{1}{M}\sum_{m=1}^{M} G^m\\ &+\frac{(e-2)\gamma^2}{K^2(1-\gamma)}\sum_{t=1}^{T}\frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M}\hat{r}_k^m(t)\end{aligned} \tag{21}$$

We know that $\log\left(\frac{W_{T+1}}{W_1}\right)$ amounts to the following:

$$\log\left(\frac{W_{T+1}}{W_1}\right) = \log(\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K}\omega_k^m(t+1)) - \log(K) \tag{22}$$

$$\geq \frac{1}{M}\sum_{m=1}^{M}\log(\sum_{k=1}^{K}\omega_k^m(t+1)) - \log(K) \tag{23}$$

$$\geq \frac{1}{M}\sum_{m=1}^{M}\log(\omega_k^m(t+1)) - \log(K) \tag{24}$$

$$\geq \frac{\gamma}{K}\frac{1}{M}\sum_{m=1}^{M}\sum_{t=1}^{T}\hat{r}_k^m(t) - \log(K) \tag{25}$$

Owing to the concavity of the logarithm function, we can go from (22) to (23). In (24), we use one term to lower bound (any channel or arm k). Finally, from (9) we can deduce that $\omega_k^m(t+1) = exp(\frac{\gamma}{K}\sum_{t=1}^{T}\hat{r}_k^m(t))$ which gives (25). Combining the two inequalities in (21) and (25) yields:

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M}G^m \geq &\frac{(1-\gamma)}{M}\sum_{m=1}^{M}\sum_{t=1}^{T}\hat{r}_k^m(t) - \frac{(1-\gamma)}{\gamma}K\log(K)\\ &-\frac{(e-2)\gamma}{K\cdot M}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{m=1}^{M}\hat{r}_k^m(t)\end{aligned} \tag{26}$$

Taking the expected value on both sides with respect to the distribution of $< k_1^m, \ldots, k_T^m >$ for all devices:

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}[G^m] \geq &\frac{(1-\gamma)}{M}\sum_{m=1}^{M}\sum_{t=1}^{T}r_k^m(t) - \frac{K}{\gamma}\log(K)\\ &-\frac{(e-2)\gamma}{K\cdot M}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{m=1}^{M}r_k^m(t)\end{aligned} \tag{27}$$

Since the channel k was chosen arbitrarily, and since

$$\sum_{t=1}^{T}\sum_{k=1}^{K}r_k^m(t) \leq K\cdot g^m$$

We obtain $\mathbb{E}[G] \geq (1-\gamma)g - \frac{K}{\gamma}\log(K) - (e-2)\gamma g$, which finally gives $g - \mathbb{E}[G] \leq \gamma(e-1)g + \frac{K}{\gamma}\log(K)$.

References

[1] Y. Zhan, J. Zhang, Z. Hong, L. Wu, P. Li, and S. Guo, "A survey of incentive mechanism design for federated learning," *IEEE Transactions on Emerging Topics in Computing*, vol. 10, no. 2, pp. 1035–1044, 2022.

[2] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[3] S. Hosseinalipour, C. G. Brinton, V. Aggarwal, H. Dai, and M. Chiang, "From federated learning to fog learning: Towards large-scale distributed machine learning in heterogeneous wireless networks," *ArXiv*, vol. abs/2006.03594, 2020.

[4] C. Shi and C. Shen, "Federated multi-armed bandits," *CoRR*, vol. abs/2101.12204, 2021.

[5] C. Shi, C. Shen, and J. Yang, "Federated multi-armed bandits with personalization," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Banerjee and K. Fukumizu, Eds., vol. 130. PMLR, 13–15 Apr 2021, pp. 2917–2925.

[6] T. Li, L. Song, and C. Fragouli, "Federated recommendation system via differential privacy," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 2592–2597.

[7] R. Huang, W. Wu, J. Yang, and C. Shen, "Federated linear contextual bandits," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[8] A. Dubey and A. Pentland, "Differentially-private federated linear bandits," *CoRR*, vol. abs/2010.11425, 2020.

[9] W. Li, Q. Song, J. Honorio, and G. Lin, "Federated x-armed bandit," *ArXiv*, vol. abs/2205.15268, 2022.

[10] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2/3, pp. 235–256, 2002.

[11] B. Sklar, "Rayleigh fading channels in mobile digital communication systems. i. characterization," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 136–146, 1997.

[12] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[13] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[14] "SimPy – event discrete simulation for python." 2018.