Adaptive NOMA in time-varying wireless networks with no CSIT/CDIT relying on a 1-bit feedback

Hajar El Hassani, Anne Savard, E Veronica Belmega

Adaptive NOMA in time-varying wireless networks with no CSIT/CDIT relying on a 1-bit feedback

Hajar El Hassani*, Anne Savard†‡, and E. Veronica Belmega*
*ETIS UMR 8051, CY Cergy Paris Université, ENSEA, CNRS, F-95000, Cergy, France
† IMT Lille Douai, Institut Mines Télécom, Centre Digital Systems, F-59653 Villeneuve d'Ascq, France
‡Univ. Lille, CNRS, Centrale Lille, UPHF, UMR 8520 - IEMN, F-59000 Lille, France
Email: hajar.el-hassani@ensea.fr, anne.savard@imt-lille-douai.fr, belmega@ensea.fr

***Abstract*—In this letter, the outage performance of a two-user downlink non-orthogonal multiple access (NOMA) time-varying network without any knowledge on the channel state or distribution at the base station (BS) is investigated. Building on reinforcement learning techniques and, more precisely, on multi-armed bandits (MAB), we propose a novel *adaptive NOMA* scheme that optimally tunes which user should perform successive interference cancellation (SIC) jointly with the power allocation at the BS. Remarkably, our proposed scheme requires only a single bit (ACK-type) of feedback from each user and is still able to outperform OMA, as demonstrated by the numerical results in many settings of interest including stochastic and even non-stationary (adversarial) ones.**

***Index Terms*—Adaptive NOMA with no CSIT/CDIT, time-varying wireless networks, reinforcement learning, multi-armed bandits, malicious jamming**

I. Introduction

Non-orthogonal multiple access (NOMA) is a promising technology for 5G and future generation communication systems aiming at solving the radio spectrum scarcity and enabling massive device connectivity [1]–[5]. Unlike orthogonal multiple access (OMA) used up to 4G, NOMA allows multiple users to share resources (frequency, time or code) by using superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver to cope with the multi-user interference.

Several studies showed that power-domain NOMA (in the uplink or downlink) can outperform OMA in terms of throughput or energy efficiency [6]–[10] when channel state information at the transmitter (CSIT) is available, as well as in terms of outage probability [11]–[14] when channel distribution information at the transmitter (CDIT) is available. Under perfect CSIT, the order under which the messages are decoded by SIC is based on the quality of the channels. When the BS does not have access to CSIT but to CDIT, NOMA can still improve both the achievable rate and the outage probability compared to OMA, by carefully allocating power to each of the served users and using a SIC ordering that is either based on the channel statistical characteristics [12], [13] or on a 1-bit feedback mechanism related to the channel state (known at the receiver end) [15].

All the above cited works studying NOMA networks assume either perfect CSIT or CDIT. Such information may be difficult to obtain at the transmitter side or even to be fed back from the receivers in practice. For instance, in dense IoT networks, when taking into account the users' heterogeneity, mobility and connectivity patterns, the network may vary too quickly to reasonably assume perfect CSIT and may vary in a completely arbitrary way (even non stationary) to assume CDIT.

To the best of our knowledge, this letter is the first to investigate NOMA networks with no CSIT and no CDIT. More precisely, we investigate a time-varying downlink NOMA network, in which a base station (BS) serves two users over wireless channels with no access to CSIT or CDIT. Our main contribution lies in the design of a novel *adaptive NOMA* scheme that jointly allocates the overall power of the BS to the two users and decide which user performs SIC decoding to minimize the network outage probability. To circumvent the lack of channel knowledge at the BS, our novel scheme exploits reinforcement learning techniques, more precisely the so-called multi-armed bandits (MAB) [16], that have a relatively low complexity and rely only on a single bit of information from each user. Our numerical simulations illustrate the enormous potential of our adaptive NOMA scheme outperforming its OMA counterpart in many settings of interest, including the presence of a malicious jammer.

MABs have recently been used in NOMA networks in [17], [18]. However, these works investigate the uplink NOMA setting, in which SIC decoding is performed at the unique receiver (having access to perfect channel state information) and is less complex compared to our downlink case. Moreover, the problems in [17], [18] are quite different and consist in throughput and energy-efficiency maximization problems (assuming perfect CSIT) respectively, as opposed to our outage minimization problem with no CSIT/CDIT.

II. System model

We consider a NOMA downlink network composed of a single-antenna transmitter, i.e., the BS (which can be an IoT access point or celullar access point etc.) and two[1] single-

This work has been supported by the ELIOT ANR-18-CE40-0030 and FAPESP 2018/12579-7 project and by IRCICA, CNRS USR 3380, Lille, France.

[1] Our analysis and results in this work carry over the more general case of multiple receivers that have been paired and assigned to orthogonal frequency bands. Pairing is a common operation in NOMA networks to reduce the complexity of the SIC decoding, which is relevant for IoT devices with limited power and computational resources [8], [19], [20].

antenna receivers (IoT or cellular users). At each time slot t, the BS transmits a signal to both users via superposition coding. The received signal at each user $k \in \{1,2\}$ at time t writes thus as $y_k^{(t)} = h_k^{(t)} \left(\sqrt{P_1^{(t)}} s_1^{(t)} + \sqrt{P_2^{(t)}} s_2^{(t)} \right) + n_k^{(t)}$.

We consider a stochastic channel model (e.g., Rayleigh fading, Rice, etc) that varies at each time instant t. The noise term $n_k^{(t)}$ follows the complex Gaussian distribution $n_k^{(t)} \sim \mathcal{CN}(0, \sigma_{n_k}^2)$, similarly to $s_k^{(t)} \sim \mathcal{CN}(0,1)$, the message intended for user k of normalized average power. Let $P_k^{(t)}$ denote the power allocated by the BS to receiver k at time t, and $P_{\max}$ denote the total power budget available at the BS, which is fully exploited: $P_1^{(t)} + P_2^{(t)} = P_{\max}$ as in [8], [12]. Also, each user k needs to meet some quality of service (QoS) requirement given as the minimum or target rate R_k^{th}.

Throughout this letter, we assume that both receivers have perfect knowledge of their own channel (which can be obtained through pilot-based channel estimation for instance) but that no CSIT nor CDIT is available at the transmitter side.

Under perfect CSIT [4], NOMA is performed as follows. The user i who encounters better channel condition $|h_i|/\sigma_{n_i}^2 > |h_j|/\sigma_{n_j}^2$, $j \in \{1,2\} \setminus \{i\}$ carries out SIC decoding, whereas the weakest user j performs single user detection (SUD). Hence, the strongest user i first detects the message of the weakest user j, cancels it out and then decodes his own signal without interference. The weakest user j decodes his own message directly by treating the interference as noise. Usually, more power is allocated by the BS to the weakest user for fairness reasons and to minimize the overall outage of the network.

When the BS does not have access to perfect CSIT, it cannot decide without error which user encounters better channel conditions, the users' decoding schemes (SIC or SUD) and its own optimal power allocation, which inevitably leads to outage events [11]–[13]. All latter works investigate the outage probability and assume perfect CDIT. Based on this information, an expression of the outage probability is provided (even in closed-form depending on the statistics of the channel) and then optimized via classical techniques.

To the best of our knowledge, our work is the first to minimize the outage probability in a downlink NOMA network with no CDIT at the BS. By exploiting reinforcement learning techniques, we propose a new adaptive NOMA scheme, in which the users' decoding choice and the power allocation at the BS are jointly tuned based on past transmissions and relying on a single bit of feedback from each user. The feedback information is of ACK-type and conveys whether the users' QoS constraints have been met during the past transmission.

In the remaining of this letter, we reserve the indices i and j to denote the user performing SIC and the user performing SUD, respectively. At each time step t, user i starts by decoding the message intended for user j, which requires the rate $R_{j\to i}^{(t)} = \log\left(1 + \Gamma_{j\to i}^{(t)}\right)$, where $\Gamma_{j\to i}^{(t)} = \frac{|h_i^{(t)}|^2 P_j^{(t)}}{|h_i^{(t)}|^2 P_i^{(t)} + \sigma_{n_i}^2}$ denotes the instantaneous signal-to-interference-plus-noise ratio (SINR) at user i. The achievable data rate of user i and j to detect their own message are respectively $R_i^{(t)} = \log(1+\Gamma_i^{(t)})$ and $R_{j\to j}^{(t)} = \log(1 + \Gamma_{j\to j}^{(t)})$, where $\Gamma_i^{(t)} = \frac{|h_i^{(t)}|^2 P_i^{(t)}}{\sigma_{n_i}^2}$ and $\Gamma_{j\to j}^{(t)} = \frac{|h_j^{(t)}|^2 P_j^{(t)}}{|h_j^{(t)}|^2 P_i^{(t)} + \sigma_{n_j}^2}$ denote the instantaneous signal-to-noise ratio (SNR) at user i after removing the interference signal, and the SINR at user j respectively. Since user i is chosen to carry out SIC, it will be allocated less power to ensure that the weakest user j does not suffer from outage too often, which would have a negative impact on the network outage. Furthermore, because the BS is assumed to transmit at full power $P_{\max}$, one can write the allocated power to user i and j respectively as $P_i^{(t)} = \alpha^{(t)} P_{\max}$ and $P_j^{(t)} = (1-\alpha^{(t)}) P_{\max}$, with $\alpha^{(t)} \in (0, 1/2)$.

To sum up, the network outage probability is defined as

$$\begin{aligned} \mathbb{P}_{\text{out}}^{\text{NOMA}} &\triangleq \mathbb{P}\Big[R_i^{(t)} \le R_i^{\text{th}} \;\cup\; \min(R_{j\to j}^{(t)}, R_{j\to i}^{(t)}) \le R_j^{\text{th}}\Big] \\ &= \mathbb{P}\Big[\Gamma_i^{(t)} \le \Gamma_i^{\text{th}} \;\cup\; \min(\Gamma_{j\to j}^{(t)}, \Gamma_{j\to i}^{(t)}) \le \Gamma_j^{\text{th}}\Big]. \quad (1) \end{aligned}$$

with $\Gamma_i^{\text{th}} \triangleq 2^{R_i^{th}} - 1$ and $\Gamma_j^{\text{th}} \triangleq 2^{R_j^{th}} - 1$.

Ideally, we would like to find the best power allocation $\alpha \in (0, 1/2)$ at the BS and the best choice $i \in \{1,2\}$ defining the user decoding schemes that minimize jointly the outage probability above. The issue is that the analytical expression of the outage probability requires the perfect knowledge of the distribution of the channel gains (or CDIT), which is unavailable in our case. Furthermore, in practical settings in which users' mobility and connectivity patterns are taken into account, the channels may vary in a completely arbitrary way and may even be non-stationary.

All this leads to an unknown objective function which cannot be minimized via classical optimization techniques. Instead, we propose to exploit reinforcement learning and MABs to propose iterative and adaptive schemes that exploit past transmissions and do not rely on CSIT/CDIT.

III. Multi-Armed Bandits for adaptive NOMA with no CSIT/CDIT

As already mentioned, we exploit here the MAB framework to design iterative policies $a^{(t)} \triangleq (\alpha^{(t)}, i^{(t)})$ that minimize the outage probability of a wireless downlink NOMA system in the absence of CSIT and CDIT. For this, we assume a discrete set: $\mathcal{Q}_\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_M\}$ of choices for the power allocation variable α. This quantization will obviously induce an optimality loss compared with the continuous set $\alpha \in (0, 1/2)$, which will be evaluated in details via numerical simulations in Sec. IV. We then denote by $\mathcal{A} = \mathcal{Q}_\alpha \times \{1,2\}$ the set of arms or policies representing the possible choices of the joint optimization variable $a \triangleq (\alpha, i)$.

A generic dynamic policy for adaptive NOMA in this framework can be described as follows and is summarized in Algorithm 1. At each iteration t, the decision maker or the BS selects an arm $a^{(t)} \in \mathcal{A}$ defining both its power allocation policy $\alpha^{(t)}$ and the users' decoding schemes $i^{(t)}$ for the transmission. Then, the BS informs the users of their decoding schemes $i^{(t)}$, which can be conveyed via 1-bit. Then, both users perform their respective decoding schemes and determine if they met their QoS requirements and send a one-bit ACK feedback. Based on this feedback, the BS computes the binary reward $u^{(t)}(a^{(t)})$ defined in (2) and updates the arm selection process.

$$u^{(t)}(a^{(t)})=\begin{cases} 1, & \text{if } \Gamma_i^{(t)} \geq \Gamma_i^{\text{th}} \cap \min(\Gamma_{j\to j}^{(t)}, \Gamma_{j\to i}^{(t)}) \geq \Gamma_j^{\text{th}} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Algorithm 1 Adaptive NOMA via MABs

```
initialize $t = 1$, $a^{(1)} = (1/4, 1)$
repeat
    1-bit broadcast of the users' decoding schemes $i^{(t)}$
    transmit data via policy $a^{(t)} = (\alpha^{(t)}, i^{(t)}) \in \mathcal{A}$
    receive a 1-bit ACK feedback from each user
    compute $u^{(t)}(a^{(t)})$ in eq. (2)
    update policy $a^{(t+1)} \leftarrow a^{(t)}$ based on $u^{(t)}(a^{(t)})$
    $t \leftarrow t + 1$
until transmission ends
```

The main idea of such iterative schemes is to learn the best policy or arm that obtains maximal expected reward, without knowing the statistics of the rewards of each arm in advance and based only on past observations. The notion of merit that allows one to evaluate the performance of a specific dynamic policy $a^{(t)}$ in the MAB framework is that of the average regret [21] defined as: $\text{Reg}_T = \mu^* - \frac{1}{T}\sum_{t=1}^{T} u^{(t)}(a^{(t)})$,where μ^* represents the maximal expected reward given as $\mu^* = \max_{a\in\mathcal{A}} \mu(a)$, with $\mu(a) = \mathbb{E}[u^{(t)}(a)]$ being the unknown expected reward of an arbitrary arm $a \in \mathcal{A}$. Intuitively, the average regret measures the gap between cumulative reward of the dynamic policy $a^{(t)}, \forall t \in \{1, \ldots, T\}$, compared with the fixed optimal policy $a^* = \arg\max_{a\in\mathcal{A}} \mu(a)$ chosen at each iteration.

A dynamic policy $a^{(t)}$ is said to have the *no-regret property* if $\limsup_{T\to\infty} \text{Reg}_T \leq 0$. This means that a no-regret dynamic policy performs at least as good as the best fixed policy maximizing the expected reward when the time horizon grows large. The link between no-regret policies and our initial problem of minimizing the outage probability is straightforward and lies in that maximizing the expected reward is equivalent to minimizing the outage probability since $\mu(a) = \mathbb{E}[u^{(t)}(a)] = 1 - \mathbb{P}_{\text{out}}^{\text{NOMA}}$.

In what follows, we focus on two well-known MAB policies that have the property of no regret by specifying the updating rule of the policy in our Algorithm 1 ($a^{(t+1)} \leftarrow a^{(t)}$). Namely, UCB and EXP3 are investigated because they optimally tradeoff between data exploration and exploitation to reach the best regret decay rates in the stochastic and adversarial MABs, respectively.

A. Upper Confidence Bound (UCB)

UCB is a deterministic no-regret algorithm designed specifically for stochastic environments which enjoys an optimal decay rate of the average regret such that $\text{Reg}_T = \mathcal{O}(\log T/T)$ [21]. Under UCB, the updating policy rule is $a^{(t+1)} = \arg\max_{a\in\mathcal{A}} \left(\hat{\mu}_a^{(t)} + \sqrt{\frac{\delta \log t}{2n_a^{(t)}}}\right)$, where $n_a^{(t)}$ denotes the number of times arm a was selected up to iteration t, $\hat{\mu}_a^{(t)}$ denotes the empirical mean reward of arm a: $\hat{\mu}_a^{(t)} = \frac{\sum_{\tau=1}^{t} u^{(\tau)}(a)\mathbb{1}[a^{(\tau)}=a]}{n_a^{(t)}}$, where $\mathbb{1}[\cdot]$ is the indicator function and δ is the learning parameter that tradeoffs between data exploration and exploitation.

B. Exponential weights for exploration and exploitation (EXP3)

EXP3 is a different and random no-regret algorithm designed for more general environments going beyond the stochastic case [22]. The regret decay rate of EXP3 is $\text{Reg}_T = \mathcal{O}(1/\sqrt{T})$, hence slower than UCB in stochastic environments, but having the advantage of accounting for arbitrary dynamics that may even be adversary as we will see in Sec. IV.

Under EXP3, the updated policy $a^{(t+1)} \in \mathcal{A}$ is drawn randomly following a discrete distribution:

$$p^{(t)}(a) = (1-\gamma)\frac{\exp(\eta G^{(t)}(a))}{\sum_{b=1}^{|\mathcal{A}|}\exp(\eta G^{(t)}(b))} + \frac{\gamma}{|\mathcal{A}|}, \quad \forall a \in \mathcal{A}$$

where $|\mathcal{A}| = 2M$ is the number of arms, $G^{(t)}(a)$ is the cumulative estimated reward of an arbitrary arm a given as $G^{(t)}(a) = \sum_{\tau=1}^{t} \hat{u}^{(\tau)}(a)\mathbb{1}[a^{\tau} = a]$, with $\hat{u}^{(t)}(a) = u^{(t)}(a)/p^{(t)}(a)$ and γ, η are the learning parameters that tradeoff between data exploration and exploitation.

Complexity of adaptive NOMA: Both UCB and EXP3 induce a linear complexity $\mathcal{O}(|\mathcal{A}|)$ per learning iteration. Also, EXP3 requires at most $\mathcal{O}(1/\epsilon^2)$ iterations to reach an average regret below $\epsilon > 0$ due to its $\mathcal{O}(1/\sqrt{T})$ regret decay rate; and since the regret decay of UCB is $\mathcal{O}(\log T/T)$ (much better than EXP3), UCB will require much less iterations compared to EXP3. Moreover, NOMA relying on perfect CSIT or CDIT involves a high-resolution feedback channel, whereas our proposed adaptive NOMA scheme via reinforcement learning requires only a single bit of feedback.

IV. Simulation results

In this section, we investigate the outage performance of our proposed adaptive NOMA schemes when no CSIT/CDIT is available at the transmitter. Four cases depending on the resolution of the power allocation interval quantization $(0, 1/2)$ are considered: a) $M = 1$ or 2 arms; b) $M = 3$ or 6 arms; c) $M = 7$ or 14 arms; and d) $M = 15$ or 30 arms. The quantization is uniform and obtained by dichotomy such that for $M = 1$ (2 arms), we have $\mathcal{A} = \{0.25\} \times \{1, 2\}$; for $M = 3$ (6 arms), we have $\mathcal{A} = \{0.125, 0.25, 0.375\} \times \{1, 2\}$; etc.

As a comparison benchmark, we consider a conventional OMA system where the BS serves both users by time sharing. The achievable rate at user $k \in \{1, 2\}$ thus write as $R_k^{(t),\text{OMA}} = \frac{1}{2}\log\left(1 + \Gamma_k^{(t),\text{OMA}}\right)$, where $\Gamma_k^{(t)} = \frac{|h_k^{(t)}|^2 P_{\max}}{\sigma_{n_k}^2}$ denotes the instantaneous SNR at user k. Note that under OMA, the outage probability writes as $\mathbb{P}_{\text{out}}^{\text{OMA}} \triangleq \mathbb{P}\left[R_1^{(t),\text{OMA}} \leq R_1^{\text{th}} \cup R_2^{(t),\text{OMA}} \leq R_2^{\text{th}}\right]$.

We evaluate our schemes in a common downlink NOMA setup [23] assuming Rayleigh channels $h_k^{(t)} \sim \mathcal{CN}(0, \sigma_{h_k}^2)$, and setting the network parameters as: $P_{\max}/\sigma_k^2 = 20$ dB for $k \in \{1, 2\}$, $\sigma_{h_1}^2 = 1, \sigma_{h_2}^2 = 0.1$, $\Gamma_1^{\text{th}} = 1$ ($R_1^{\text{th}} = 1$ bpcu), $\Gamma_2^{\text{th}} = 3$ ($R_2^{\text{th}} = 2$ bpcu), unless otherwise specified. Both algorithms are run over a $T = 10^4$ time horizon and the provided results are averaged over $N = 10^3$ random runs.

Tunning the learning parameters: δ, γ and η that tradeoff between data exploration and exploitation can lead to very poor performance if badly chosen. From a theoretical perspective, they should be chosen such that $\delta^* > 2$, $\eta^* = \frac{\gamma}{|\mathcal{A}|}$ and $\gamma^* = \sqrt{\frac{|\mathcal{A}|\ \ln|\mathcal{A}|}{(e-1)T}}$ in order to reach the optimal regret decay rate by minimizing the regret's upper bound [21], [22]. In practice, these values can be further improved to obtain better performance [24]–[26]. Based on numerical experiments, the following values were chosen: $\delta = 1$ instead of $\delta^* > 2$ (for UCB in Fig. 1 and Fig. 3), $\gamma = \gamma^* = 0.0464$ and $\eta = 0.02$ instead of $\eta^* = 0.0033$ (for EXP3 in Fig. 1).

In Fig. 1, we compare the outage performance obtained with UCB and EXP3 using a set of 14 arms ($M = 7$) with the fixed optimal arm a^* (computed offline and requiring CDIT) using a set of 14 arms ($M = 7$) as well as with OMA. Notice that both algorithms converge towards a^*, the best offline solution by exploiting only 1-bit of feedback from the users as opposed to perfect CSIT or CDIT. Surprisingly, our proposed adaptive NOMA schemes quickly outperform OMA (after less than 100 iterations). Finally, UCB performs better than EXP3 as expected in stochastic environments.

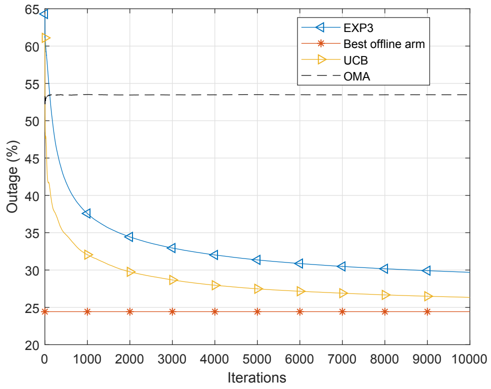

Fig. 1. Outage of adaptive NOMA (via UCB or EXP3) relying on a 1-bit feedback compared to OMA and the best offline policy. Our schemes greatly outperforms OMA in terms of outage probability.

Fig. 2 depicts the outage performance of our adaptive NOMA schemes for different number of arms and $M \in \{1, 3, 7, 15\}$ as a function of the QoS requirement Γ_2^{th} of user 2. Here, aside from the OMA benchmark, we also include the optimal outage probability obtained over a continuous power allocation policy $\alpha \in (0, 1/2)$, to assess the optimality loss of our adaptive NOMA schemes based on quantization.

First, we remark that our adaptive NOMA schemes cannot decrease the outage performance compared to OMA and that the gap between both access techniques is maximized for moderate QoS requirement, and can go up to 48% in the case of 30 arms. Also, we can see that increasing the number of arms, which increases the resolution of our power allocation quantization, allows to reduce the gap with the continuous optimal NOMA transmission scheme. At last, for low QoS requirements two arms are sufficient for outage optimality; however, as the QoS requirement increases, the number of arms has to increase.

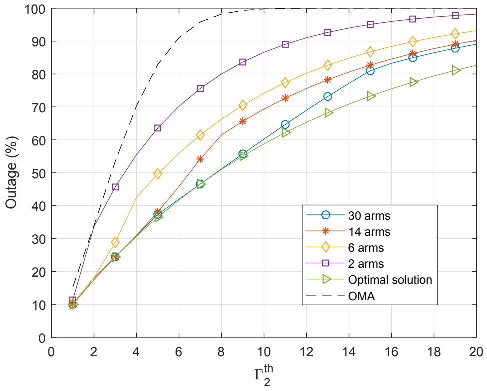

Fig. 2. Impact of the number of arms on the outage probability. The outage decays with the number of arms. The quantization incurred optimality loss becomes negligible when choosing carefully the number of arms.

So far, we have compared NOMA and OMA from an outage probability perspective. Since our adaptive NOMA approach relies on an iterative procedure, we also consider here the convergence speed. In Fig. 3, we compare the number of iterations required for adaptive NOMA with UCB to reach a regret of 10% for $M \in \{1, 3, 7, 15\}$. When the number of arms is increased, the longer it takes until UCB reaches the 10% level of regret. This can be explained by the fact that, when increasing the number of possible arms or policies, the duration of the exploration search for the best arm naturally increases.

To sum up, Fig. 1 and Fig. 3 highlight an important tradeoff between outage optimality and latency of our adaptive NOMA schemes based on MAB. Indeed, the number of arms needs to be large enough to reduce the optimality loss caused by our quantization, but not too large to insure fast convergence. Hence, the best tradeoff and number of arms will depend on the specific application.

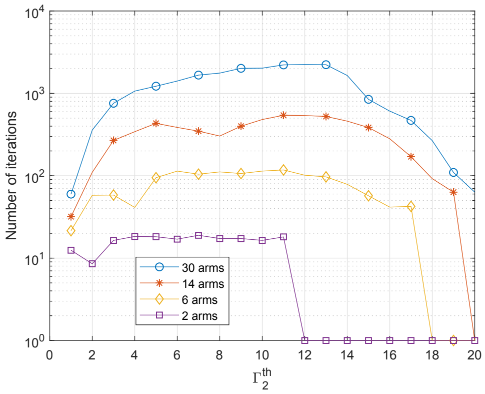

Fig. 3. Impact of the number of arms on the speed of the adaptive NOMA scheme with UCB (iterations required to reach 10% regret). The exploration search increases with the number of arms.

Let us now consider the presence of a malicious jammer whose aim is to put the network systematically in outage. For simplicity, we assume $\mathcal{A} = \{0.4\} \times \{1, 2\}$ (only two arms) and set $\delta = 1$ for UCB, $\gamma = \gamma^* = 0.009$, $\eta = 0.02$ for EXP3. The jammer is assumed to have knowledge of the network, more precisely, it knows the set of actions and the adaptive NOMA algorithm used at the base station. Since UCB is a purely deterministic algorithm, the jammer can anticipate precisely the arm or action chosen by the base station and is able to adjust its jamming power such that the network is systematically put in outage. This leads to a non-vanishing linear regret. The jammer cannot impact the network to such an extent when the base station is using EXP3. Indeed, with EXP3 the arm is randomly chosen following a probability distribution and cannot be perfectly anticipated even is such worst-case adversarial settings.

The outage performance of our adaptive NOMA schemes in the presence of a malicious jammer are depicted in Fig. 4. We can see that UCB is always in outage, as expected; and that EXP3 can still reach the best offline policy.

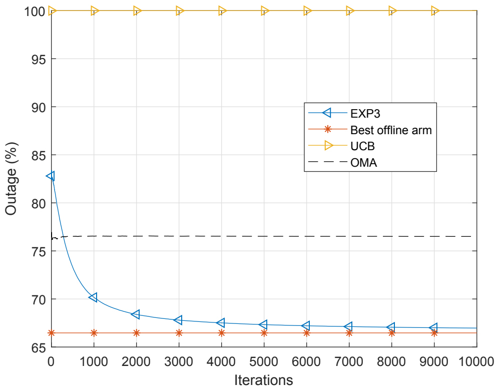

Fig. 4. Outage of adaptive NOMA (via UCB or EXP3) relying on a 1-bit feedback compared to OMA and the best offline policy in the presence of a malicious jammer. UCB is always in outage, while EXP3 outperforms OMA and reaches the best offline policy.

V. Conclusions and Perspectives

In this letter, we investigated the outage probability of a two-user downlink NOMA network in which no channel state or distribution information is available at the transmitter side. To overcome the lack of information, we exploit the multi-armed bandit (MAB) framework and propose a novel adaptive NOMA scheme relying on a single bit of feedback. In our scheme, the transmitter decides both the decoding schemes of the two receivers as well as its allocated transmit power based on two well-known reinforcement learning algorithms, namely UCB and EXP3. Our numerical results demonstrate the enormous potential of our adaptive NOMA scheme relying on 1-bit feedback by outperforming OMA in many settings of interest including stochastic and even non-stationary (adversarial) ones. Furthermore, our simulations show that the number of possible arms needs to be chosen sufficiently large to compensate for the power allocation quantization, but not too large to allow a fast outage decay.

Interesting future work includes taking into account user mobility and heterogeneous connectivity patterns, multiple users with and without pairing.

References

[1] M. Aldababsa, M. Toka, S. Gökçeli, G. K. Kurt, and O. Kucur, “A tutorial on nonorthogonal multiple access for 5G and beyond,” *Wireless Commun. and Mobile Comp.*, vol. 2018, 2018.

[2] S. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, “Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges,” *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2016.

[3] M. Vaezi, Z. Ding, and H. V. Poor, *Multiple access techniques for 5G wireless networks and beyond*, 2019.

[4] K. Higuchi and A. Benjebbour, “Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access,” *IEICE Trans. on Commun.*, vol. 98, no. 3, pp. 403–414, 2015.

[5] M. Pischella, I. Stupia, and L. Vandendorpe, “Performance analysis of uplink adaptive NOMA depending on channel knowledge,” *arXiv preprint arXiv:2004.02630*, 2020.

[6] Z. Chen, Z. Ding, X. Dai, and R. Zhang, “An optimization perspective of the superiority of NOMA compared to conventional OMA,” *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 5191–5202, 2017.

[7] Z. Yang, W. Xu, C. Pan, Y. Pan, and M. Chen, “On the optimality of power allocation for NOMA downlinks with individual QoS constraints,” *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1649–1652, 2017.

[8] L. Zhu, J. Zhang, Z. Xiao, X. Cao, and D. O. Wu, “Optimal user pairing for downlink non-orthogonal multiple access (NOMA),” *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 328–331, 2018.

[9] W. U. Khan, F. Jameel, T. Ristaniemi, S. Khan, G. A. S. Sidhu, and J. Liu, “Joint spectral and energy efficiency optimization for downlink NOMA networks,” *IEEE Trans. on Cogn. Commun. Netw.*, 2019.

[10] H. El Hassani, A. Savard, and E. V. Belmega, “A closed-form solution for energy-efficiency optimization in multi-user downlink NOMA,” in *IEEE PIMRC*, 2020.

[11] J. Cui, Z. Ding, and P. Fan, “A novel power allocation scheme under outage constraints in NOMA systems,” *IEEE Signal Process. Lett.*, vol. 23, no. 9, pp. 1226–1230, 2016.

[12] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, “On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users,” *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, 2014.

[13] X. Wang, J. Wang, L. He, and J. Song, “Outage analysis for downlink NOMA with statistical channel state information,” *IEEE Wireless Commun. Lett.*, vol. 7, no. 2, pp. 142–145, 2017.

[14] D. Tweed, M. Derakhshani, S. Parsaeefard, and T. Le-Ngoc, “Outage-constrained resource allocation in uplink NOMA for critical applications,” *IEEE Access*, vol. 5, pp. 27 636–27 648, 2017.

[15] P. Xu, Y. Yuan, Z. Ding, X. Dai, and R. Schober, “On the outage performance of non-orthogonal multiple access with 1-bit feedback,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6716–6730, 2016.

[16] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[17] M. A. Adjif, O. Habachi, and J.-P. Cances, “Joint channel selection and power control for NOMA: A multi-armed bandit approach,” in *IEEE WCNCW*, 2019, pp. 1–6.

[18] Z. Tian, J. Wang, J. Wang, and J. Song, “Distributed NOMA-based multi-armed bandit approach for channel access in cognitive radio networks,” *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1112–1115, 2019.

[19] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, “Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges,” *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 176–183, 2017.

[20] M. Vaezi, R. Schober, Z. Ding, and H. V. Poor, “Non-orthogonal multiple access: Common myths and critical questions,” *IEEE Trans. Wireless Commun.*, vol. 26, no. 5, pp. 174–180, 2019.

[21] S. Bubeck and N. Cesa-Bianchi, “Regret analysis of stochastic and nonstochastic multi-armed bandit problems,” *arXiv preprint arXiv:1204.5721*, 2012.

[22] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, “Gambling in a rigged casino: The adversarial multi-armed bandit problem,” in *IEEE 36th Annual Foundations of Computer Science*, 1995, pp. 322–331.

[23] Q. Zhang, L. Zhang, Y.-C. Liang, and P.-Y. Kam, “Backscatter-NOMA: A symbiotic system of cellular and Internet-of-Things networks,” *IEEE Access*, vol. 7, pp. 20 000–20 013, 2019.

[24] 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.

[25] P. Perick, D. L. St-Pierre, F. Maes, and D. Ernst, “Comparison of different selection strategies in monte-carlo tree search for the game of tron,” in *IEEE CIG*, 2012, pp. 242–249.

[26] A. Garivier and E. Moulines, “On upper-confidence bound policies for non-stationary bandit problems,” *arXiv preprint arXiv:0805.3415*, 2008.