Energy-efficient 1-bit feedback NOMA in wireless networks with no CSIT/CDIT

Hajar El Hassani, Anne Savard, E Veronica Belmega

ENERGY-EFFICIENT 1-BIT FEEDBACK NOMA IN WIRELESS NETWORKS WITH NO CSIT/CDIT

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 for 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 paper, the energy efficiency of a two-user downlink non-orthogonal multiple access (NOMA) system is investigated. We consider a stochastic time-varying channel that is unknown at the transmitter side (in state and distribution) and propose an online learning method that is energy efficient and relies only on a 1-bit feedback, of high relevance in IoT networks. Our adaptive NOMA scheme draws on multi armed bandits (MAB) to learn the optimal policy in terms of which user should perform successive interference cancellation (SIC) and the power allocation at the transmitter. Our simulation results reveal fundamental tradeoffs between performance, complexity and available feedback information, and also show that our scheme can remarkably outperform OMA without CSIT/CDIT.

***Index Terms*—** Adaptive NOMA with no CSIT/CDIT, energy efficiency, dynamic IoT networks, multi-armed bandits

1. INTRODUCTION

Non-orthogonal multiple access (NOMA) is a very promising technology for IoT and future generation networks due to its massive connectivity and spectrum efficiency potential [1]. In addition, IoT devices will have to cope with the network dynamics under stringent power and computational constraints [2, 3]. Hence, low-complex and energy-efficient learning capabilities relying on machine learning need to be integrated within the network's architecture in order to overcome these challenges.

The majority of works investigating NOMA [4–10] rely on strong assumptions regarding the channel knowledge at the transmitter side: either perfect channel state information at the transmitter (CSIT) or channel distribution information at the transmitter (CDIT). Under such assumptions, the order under which the messages are decoded by SIC is based on the quality of channels (either the ordering of the channel gains, or some channel statistical characteristics) and by carefully allocating the transmit powers, NOMA is shown to outperform OMA.

Assuming perfect CSIT or CDIT might not be realistic because of the extensive overhead and computational cost of CSI/CDI estimation and feedback, which is especially problematic in IoT networks composed of low-cost devices with low power and computational capabilities. In our previous work [11], we proposed an adaptive NOMA scheme exploiting reinforcement learning and relying only on a single bit of feedback that minimizes the outage probability in a time-varying two-user downlink network with no CSIT/CDIT.

In this paper, we focus on the energy efficiency maximization in a stochastic time-varying two-user downlink NOMA network with no CSIT/CDIT. Although the probability of outage is an important measure, energy efficiency becomes crucial especially for IoT networks with limited battery lifetime or low-power capability devices. We propose to use the multi-armed bandits (MABs) toolbox allowing to cope with the network dynamics and leading to low-complexity NOMA schemes relying on a 1-bit feedback mechanism, relevant to IoT networks.

Although we exploit the MAB-toolbox as in [11], the problem under study in this work is completely different. First, as opposed to minimizing the outage probability, we focus on maximizing the energy efficiency defined as the ratio between the long-term average goodput (i.e., amount of successfully transmitted information, which meets the QoS minimum rate constraints) and the power consumption. For this different objective, transmitting at full power is no longer optimal as in [11]. This leads to a different feasible set than in [11], which adds a significant challenge to our MAB formulation: our power allocation policy has to be quantized, leading to a non-trivial tradeoff between performance and complexity.

Our numerical results show that our energy-efficient adaptive NOMA scheme can outperform its OMA counterpart with no CSIT/CDIT. Moreover, we evaluate the compromise between performance vs. complexity along with the compro-

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

mise between performance and available feedback.

2. SYSTEM MODEL

We consider a downlink NOMA wireless system composed of one base station (BS) or access point and two users or IoT nodes that are multiplexed on the same resource block. The channels are assumed to follow a stochastic time-varying small-scale fading model as in [12]. Regarding the available information, we assume that only the receivers know their own channels, whereas perfect CSIT or CDIT is not available at the transmitter side. At each time instant t, the transmitter broadcasts the two-user superimposed message $x^{(t)} = \sqrt{p_1^{(t)}}x_1^{(t)} + \sqrt{p_2^{(t)}}x_2^{(t)}$, where $x_k^{(t)} \sim \mathcal{CN}(0,1),\ k \in \{1,2\}$ denotes the instantaneous message intended for user k and p_k is its allocated power.

The received signal at each user k writes as $y_k^{(t)} = h_k^{(t)}x^{(t)} + z_k^{(t)}$, where $h_k^{(t)}$ represents the instantaneous channel gain unknown to the transmitter and $z_k^{(t)} \sim \mathcal{CN}(0,\sigma^2)$ is the additive white Gaussian noise at user k. We further denote by $P_{\max}$ the power budget of the BS or access point such that $p_1 + p_2 \leq P_{\max}$. Also, each user k is required to meet a quality of service (QoS) constraint expressed in terms of a minimum achievable rate of R_k^{th} bps.

We discuss below two particular cases before delving into the stochastic channels case with no CSIT/CDIT.

A. Static and known channels at the transmitter: the user with the best channel conditions carries out successive interference cancellation (SIC) and the other performs single user detection (SUD). Assuming equal QoS constraints, the user performing SIC is allocated less power than the other user at the optimal solution. The optimal power allocation policy maximizing the energy efficiency defined as the tradeoff between the sum rate and power consumption has been provided in closed form in [6].

B. Stochastic channels and CDIT: the channels are not known so the decoding techniques of the users have to be chosen differently. One solution is to consider it as an optimization variable. More precisely, we use indices $i \in \{1,2\}$ and $j \in \{1,2\} \setminus \{i\}$ to denote the user performing SIC and the one performing SUD, respectively. The discrete variable $i \in \{1,2\}$ assigning the users' decoding techniques is a control variable that has to be optimized at the transmitter alongside the power allocation vector $\mathbf{p} = (p_i, p_j)$.

Following the NOMA principle, user i first decodes the message of user j with the data rate $R_{j\to i}^{(t)} = \log\left(1 + \Gamma_{j\to i}^{(t)}\right)$, where $\Gamma_{j\to i}^{(t)} = \frac{|h_i^{(t)}|^2 p_j}{|h_i^{(t)}|^2 p_i + \sigma^2}$ denotes the instantaneous signal-to-interference-plus-noise ratio (SINR) at user i when decoding user j's message. User i then removes this message and decodes its own message with the data rate $R_i^{(t)} = \log(1 + \Gamma_i^{(t)})$, where $\Gamma_i^{(t)} = \frac{|h_i^{(t)}|^2 p_i}{\sigma^2}$ denotes the signal-to-noise ratio (SNR) at user i. User j treats user i's message as additional noise, with the data rate $R_{j\to j}^{(t)} = \log(1 + \Gamma_{j\to j}^{(t)})$, where $\Gamma_{j\to j}^{(t)} = \frac{|h_j^{(t)}|^2 p_j}{|h_j^{(t)}|^2 p_i + \sigma^2}$ is the signal to interference plus noise ratio (SINR) at user j.

An energy efficient measure in the stochastic small-scale fading channels is defined as follows [13, 14]

$$GEE(i,\mathbf{p}) = \frac{(R_1^{\text{th}} + R_2^{\text{th}})(1 - \mathbb{P}_{\text{out}}(i,\mathbf{p}))}{p_i + p_j + P_c}, \qquad (1)$$

where P_c denotes the circuit power and $(1 - \mathbb{P}_{\text{out}})$ is the success probability or the probability that the QoS constraints are met. On the opposite side, the outage probability is $\mathbb{P}_{\text{out}}(i,\mathbf{p}) = \mathbb{P}\left[\Gamma_i^{(t)} \leq \Gamma_i^{\text{th}} \cup \min(\Gamma_{j\to j}^{(t)}, \Gamma_{j\to i}^{(t)}) \leq \Gamma_j^{\text{th}}\right]$, with $\Gamma_i^{\text{th}} \triangleq 2^{R_i^{\text{th}}} - 1$ and $\Gamma_j^{\text{th}} \triangleq 2^{R_j^{\text{th}}} - 1$.

The energy efficiency measure in (1) is relevant in small-scale fading channels as the numerator $(R_1^{\text{th}} + R_2^{\text{th}})(1 - \mathbb{P}_{\text{out}})$ represents the long-term average sum rate of the system. Also, GEE incorporates the QoS constraint in the objective, simplifying the feasible set of the problem as follows: $\mathcal{P} = \{(i,\mathbf{p}) \mid i \in \{1,2\}, p_i \geq 0, p_j \geq 0, p_i + p_j \leq P_{\max}\}$.

When CDIT is available, the closed-form expressions of the outage probability are known [9] for Rayleigh distributed channels. Hence finding the optimal policy maximizing $GEE(i,\mathbf{p})$ reduces to solving two continuous optimization problems (one for each value of i) with respect to $\mathbf{p}$ and then choosing the best value of i.

In the absence of CDIT, we propose here to exploit multi-armed bandits (MABs) similarly to [11]. Nevertheless, the problem under study is entirely different: we maximize the energy efficiency as opposed to minimizing the outage probability; also, transmitting at full power is no longer optimal here leading to a different feasible set.

3. MULTI-ARMED BANDITS FOR ENERGY EFFICIENT NOMA WITH NO CSIT/CDIT

In order to exploit the MAB framework, we first need to quantize the feasible set $\mathcal{P}$. Given that transmitting at full power as in [11] is not energy-efficient in general, we consider that only a fraction of the maximum budget $P_{\max}$ is exploited with $\beta \in \mathcal{B} \subset [0,1]$ such that $\mathcal{B} = \{\beta_1, \beta_2, \ldots, \beta_B\}$ is discrete. In order to maintain fairness among users, user i carrying out SIC is allocated less power than user j; and focus on the special choice of power allocation policy $\mathbf{p}_\beta = (0.25\beta P_{\max}, 0.75\beta P_{\max})$. Of course, the quantization set $\mathcal{B}$ and the $0.25 - 0.75$ power split between the two users will both incur an optimality loss which will be evaluated and analysed throughly via numerical simulations.

A possible action or arm at the BS is defined by the pair $\mathbf{a} \triangleq (i,\beta) \in \mathcal{A} = \{1,2\} \times \mathcal{B}$ which will dictate both the decoding schemes of the two users via i and the discrete transmit power allocation policy $\mathbf{p}_\beta$ via β as described above.

At last, another important ingredient is the energy-efficient instantaneous reward:

$$u^{(t)}(\mathbf{a})=\begin{cases}\frac{R_1^{\rm th}+R_2^{\rm th}}{p_{i,\beta}+p_{j,\beta}+P_c}, & \text{if } \Gamma_i^{(t)}\geq\Gamma_i^{\rm th} \text{ and} \\ & \min(\Gamma_{j\to j}^{(t)},\Gamma_{j\to i}^{(t)})\geq\Gamma_j^{\rm th}, \\ 0, & \text{otherwise,}\end{cases} \quad (2)$$

chosen such that its expectation equals precisely the energy efficiency measure in (1), $\mathbb{E}[u^{(t)}(\mathbf{a})] = GEE(i,\mathbf{p}_\beta)$. Therefore, maximizing the energy efficiency amounts to maximizing the expectation of the reward $\mu^* \triangleq \max_{\mathbf{a}\in\mathcal{A}} \mathbb{E}[u^{(t)}(\mathbf{a})] \equiv \max_{i\in\{1,2\},\beta\in\mathcal{B}} GEE(i,\mathbf{p}_\beta)$.

We are now ready to describe our online energy-efficient NOMA approach: at time instant t, the BS selects an arm $\mathbf{a}^{(t)} = (i^{(t)},\beta^{(t)})$. Then it informs both users of their decoding techniques dictated by $i^{(t)}$ via 1-bit broadcast (e.g., 1 if $i^{(t)}=1$ and 0 otherwise) and it transmits the superimposed signal using the $\mathbf{p}_\beta^{(t)}$ power allocation policy described above. After the transmission, each user decodes his message following the policy assigned by the BS (either SIC or SUD) and feeds back a single bit indicating whether his QoS requirement was met or not. Based on the two feedback bits, the BS computes its reward defined in (2) and updates the choice of the next arm $\mathbf{a}^{(t+1)}$. This online process is summarized in Algorithm 1.

Algorithm 1 MAB for NOMA with no CSIT or CDIT

Initialize: $t=1$, $\mathbf{a}^{(1)}=(1,0.5)$ arbitrarily
repeat
• choose policy $\mathbf{a}^{(t)} = (i^{(t)},\beta^{(t)})$: inform users of their decoding schemes ($i^{(t)}$) and transmit with power policy $\mathbf{p}_\beta^{(t)} = (0.25\,\beta^{(t)}\,P_{\max},\ 0.75\,\beta^{(t)}\,P_{\max})$
• receive 1-bit ACK feedback from each user
• compute reward $u^{(t)}(\mathbf{a}^{(t)})$ given in (2)
• update policy $\mathbf{a}^{(t+1)} \leftarrow \mathbf{a}^{(t)}$ via UCB or EXP3
• $t \leftarrow t+1$
until end of transmission

The performance of such an online learning algorithm is measured by the *regret* [15] defined as: $\mathrm{Reg}(T) = T\mu^* - \sum_{t=1}^{T} u^{(t)}(\mathbf{a}^{(t)})$, which represents the gap between the optimal energy efficiency $GEE(i,\mathbf{p}_\beta)$ and the overall performance of the online algorithm $\mathbf{a}(t)$ over a fixed horizon of time T. A desirable property is that of no regret, i.e., $\limsup_{T\to\infty} \mathrm{Reg}(T) \leq 0$, which implies that the online algorithm has to perform at least as good as the optimal arm maximizing the energy efficiency.

In the following, we analyze two among the most popular MAB algorithms, namely UCB and EXP3[16–18], that have the property of no regret in the stochastic case. More precisely, by optimally tuning the learning parameters which tradeoff between past reward exploitation and arm exploration, the regret of UCB and EXP3 decays to zero as $\mathcal{O}(\log T/T)$ and $\mathcal{O}(\sqrt{T})$ [16–18], respectively. Regarding their complexity, both are fairly simple algorithms as each update or iteration scales linearly with the number of arms: $\mathcal{O}(|\mathcal{A}|)$.

For the sake of completeness, we briefly present the updating rules for both UCB and EXP3, but kindly refer the interested reader to [11] for details.

Under UCB, the updating rule is deterministic and given as $\mathbf{a}^{(t+1)} = \arg\max_{\mathbf{a}\in\mathcal{A}} \left(\hat{\mu}_{\mathbf{a}}^{(t)} + \sqrt{\frac{\alpha\log t}{2n_{\mathbf{a}}^{(t)}}}\right)$, where $n_{\mathbf{a}}^{(t)}$ is the number of times arm a was selected up to iteration t, $\hat{\mu}_{\mathbf{a}}^{(t)}$ denotes the empirical mean reward of arm a up to time t and α is the learning parameter.

Under EXP3, the arm $\mathbf{a}^{(t+1)} \in \mathcal{A}$ is drawn randomly following the updated probability distribution $q^{(t+1)}(\mathbf{a}) = \frac{q^{(t)}(\mathbf{a})\exp(\eta\ \hat{u}^{(t)}(\mathbf{a}))}{\sum_{b=1}^{|\mathcal{A}|} q^{(t)}(\mathbf{b})\exp(\eta\ \hat{u}^{(t)}(\mathbf{b}))}$, $\forall \mathbf{a} \in \mathcal{A}$, where $\hat{u}^{(t)}(\mathbf{b}) = u^{(t)}(\mathbf{a}^{(t)})\ \mathbb{1}_{[\mathbf{b}=\mathbf{a}^{(t)}]}/q^{(t)}(\mathbf{b})$ is the estimated reward of arm $\mathbf{b}$ and η is the learning parameter.

Notice that our low-power feedback mechanism is especially relevant for IoT networks connecting low-power wireless sensors. Moreover, via this simple online process, the BS is capable of learning the best decoding scheme (SIC/SUD) for each user and the best power allocation in the discretized set maximizing the energy efficiency in (1), without requiring CSIT nor CDIT and only a 1-bit feedback.

4. SIMULATION RESULTS

Let us now investigate the energy efficiency of our proposed adaptive NOMA scheme with no CSIT/CDIT. Three cases based on the quantization set $\mathcal{B}$ are considered: a) 5-element $\mathcal{B}_1 = \{0.2, 0.4, \ldots 1\}$; b) 10-element $\mathcal{B}_2 = \{0.1, 0.2, \ldots, 1\}$; and c) 20-element $\mathcal{B}_3 = \{0.05, 0.1, \ldots, 1\}$ such that $\mathcal{B}_1 \subset \mathcal{B}_2 \subset \mathcal{B}_3$. The considered arm sets are thus given as $\mathcal{A}_i = \{1,2\} \times \mathcal{B}_i$ of 10, 20 and 40 arms respectively.

Our adaptive scheme is evaluated in a downlink NOMA setup where we assume Rayleigh fading channels, i.e, $h_k^{(t)} \sim \mathcal{CN}(0,1)$, and consider the following system parameters: $\sigma^2 = 0.1$, $\Gamma_1^{\rm th} = 1$ ($R_1^{\rm th} = 1$ bpcu), $\Gamma_2^{\rm th} = 10$ ($R_2^{\rm th} \simeq 3.5$ bpcu), $P_c = 1$ W and $P_{\max} = 100$ W (unless stated otherwise). The time horizon is set to $T = 5000$ for both UCB and EXP3 algorithms and the illustrated curves are averaged over 10^3 horizon realizations. The learning parameters were carefully tuned in order to provide the best performance and were set as $\alpha = 0.1$ and $\eta = 0.08$.

To benchmark the performance of our proposed adaptive NOMA scheme, we consider OMA where each of the two users is served in a time-sharing manner. The achievable rate and energy efficiency of user k write as $R_k^{(t),\rm OMA} = \frac{1}{2}\log\left(1+\frac{|h_k^{(t)}|^2\beta P_{\max}}{\sigma^2}\right)$ and $GEE^{\rm OMA}(\beta) = \frac{(R_1^{\rm th}+R_2^{\rm th})(1-\mathbb{P}_{\rm out}^{\rm OMA}(\beta))}{\beta P_{max}+P_c}$. The optimal $\beta^{*,\rm OMA}$ is obtained offline with the help of CDIT.

In Fig. 1, we compare the energy efficiency of our NOMA scheme with UCB and EXP3 using $\mathcal{A}_3$ (40 arms), with the fixed optimal arm a^* computed offline with the use of CDIT. Note that both algorithms reach the energy efficiency of a^*,

the best fixed offline policy, by requiring only one bit of feedback and no CSIT/CDIT. Further, our proposed NOMA scheme significantly outperforms OMA after a few iterations. UCB outperforms EXP3 in the stochastic case [18].

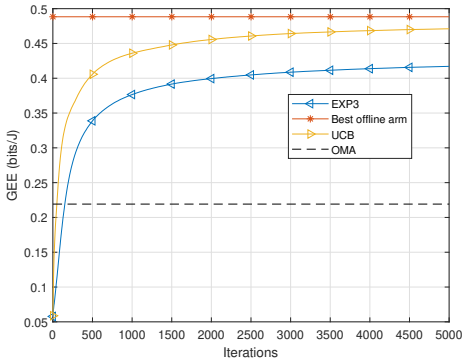

Fig. 1. Energy efficiency of our adaptive NOMA (via UCB or EXP3) compared to the best offline arm and OMA.

In Fig. 2 and Fig. 3, we investigate the impact of the number of arms and the sub-optimality caused by the discretization $\mathcal{B}$ and the split $0.25-0.75$ for two scenarios: $P_{\max} = 100$ W and $P_{\max} = 10$ W, respectively. For this, we include the following benchmarks: a) sub-optimal energy efficiency obtained with the user power split $0.25-0.75$, but for an optimal choice of β; b) the optimal energy efficiency obtained over the entire set $\mathcal{P}$.

Both figures show a vanishing gap between the above sub-optimal and optimal schemes, showing hence the efficiency of our heuristic $0.25-0.75$ power split between the two users. In the large transmit power regime of Fig. 2, the optimality loss of our adaptive NOMA scheme decreases with the number of arms. However, the gap remains large (more than 50% at low Γ_2^{th}), highlighting the *tradeoff between energy efficiency and available feedback*. On the other hand, in the low power regime of Fig. 3, the optimality gap is negligible for 20 arms.

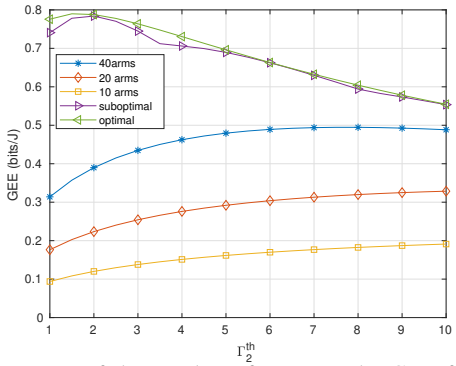

Fig. 2. Impact of the number of arms on the GEE for $P_{\max} = 100$ W: tradeoff performance vs. available information.

In Fig. 4, we study how the number of arms affects the regret performance of our adaptive NOMA. We focus only on UCB, since it is known to have a better decay rate than EXP3 in the stochastic case, and plot the number of iterations required to reach a regret level of 10%. We see that the larger the number of arms, the more iterations are needed. Hence, even if a better energy efficiency performance can be achieved by increasing the number of arms, additional time is needed to explore and exploit all arms. This highlights another fundamental *tradeoff between energy efficiency and complexity* and a larger amount of time is required to reach better performance.

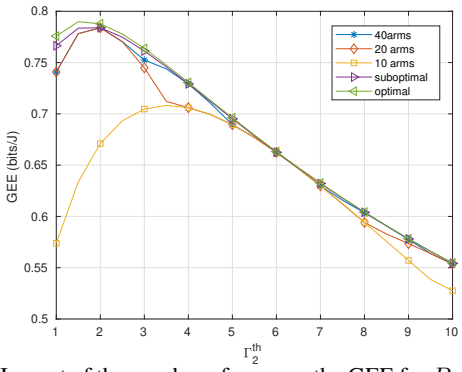

Fig. 3. Impact of the number of arms on the GEE for $P_{\max} = 10$ W.

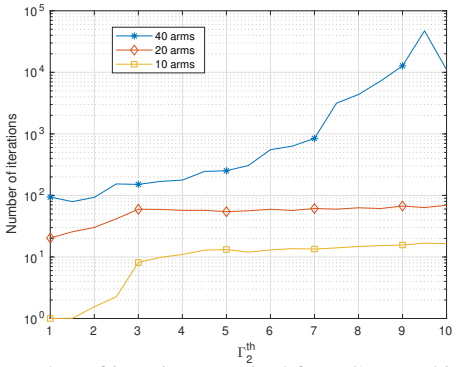

Fig. 4. Number of iterations required for UCB to achieve 10% regret level: tradeoff performance vs. complexity.

5. CONCLUSIONS AND PERSPECTIVES

In this paper, we maximize the energy efficiency of a dynamic two-user downlink NOMA system with no CSIT/CDIT. Exploiting multi-armed bandits (MABs) and the well-known UCB and EXP3 algorithms, we propose an adaptive enegy-efficient NOMA scheme relying only on a 1-bit feedback, relevant for IoT networks. Our simulation results show that our adaptive NOMA scheme can outperform OMA in stochastic environments with no CSIT/CDIT. The fundamental tradeoffs between performance, feedback information and complexity are highlighted, indicating that the number of arms for MABs needs to be carefully tuned depending on the specific application requirements and constraints.

Future works may include non stationary channels (e.g, adversarial environments), in which EXP3 is expected to outperform UCB, and continuous sets of arms to reduce the optimality gap, etc.

6. REFERENCES

[1] B. Makki, K. Chitti, A. Behravan, and M. S. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 179–189, 2020.

[2] I. Ahmad, S. Shahabuddin, T. Sauter, E. Harjula, T. Kumar, M. Meisel, M. Juntti, and M. Ylianttila, "Challenges of AI in wireless networks for IoT," *IEEE Ind. Electron. Mag.*, pp. 0–0, 2020.

[3] W. U. Khan, F. Jameel, M. A. Jamshed, H. Pervaiz, S. Khan, and J. Liu, "Efficient power allocation for NOMA-enabled IoT networks in 6G era," *Physical Communication*, vol. 39, p. 101043, 2020.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] H. El Hassani, A. Savard, and E. V. Belmega, "Adaptive NOMA in time-varying wireless networks with no CSIT/CDIT relying on a 1-bit feedback," *IEEE Wireless Commun. Lett.*, 2020.

[12] G. Lee, M. Jung, A. T. Z. Kasgari, W. Saad, and M. Bennis, "Deep reinforcement learning for energy-efficient networking with reconfigurable intelligent surfaces," in *IEEE ICC 2020*. IEEE, 2020, pp. 1–6.

[13] H. Zhang, J. Zhang, and K. Long, "Energy efficiency optimization for NOMA UAV network with imperfect CSI," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 12, pp. 2798–2809, 2020.

[14] D. Ghosh, M. K. Hanawal, and N. Zlatanov, "Learning to optimize energy efficiency in energy harvesting wireless sensor networks," *arXiv preprint arXiv:2012.15203*, 2020.

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

[16] 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.

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

[18] E. Belmega, P. Mertikopoulos, R. Negrel, and L. Sanguinetti, "Online convex optimization and no-regret learning: Algorithms, guarantees and applications. 2018," *URL http://arxiv. org/abs/1804*, vol. 4529, 1804.