Max-Min Fairness with Selection Combining Strategy on Cooperative NOMA: A Finite Blocklength Analysis

Fateme Salehi*, Naaser Neda*, Mohammad-Hassan Majidi*, Hamed Ahmadi†
*Faculty of Electrical and Computer Engineering, University of Birjand, Iran
†Department of Electronic Engineering, University of York, United Kingdom

***Abstract*—In this paper, the performance of a cooperative relaying technique in a non-orthogonal multiple access (NOMA) system is considered in short packet communications with finite blocklength (FBL) codes. We examine the performance of a decode-and-forward (DF) relaying along with selection combining (SC) strategy at the receiver. Our goal is optimal power and blocklength allocation to attain maximum users' fair throughput in a downlink NOMA (DL-NOMA) system with two users, where the user with a stronger channel (strong user) acts as a relay for the user with a weaker channel (weak user). For this purpose, an optimization problem is formulated and an analytical solution is proposed. Numerical results show that the proposed Cooperative NOMA scheme improves the users' fair throughput in comparison with the NOMA in the FBL regime.**

***Index Terms*—finite blocklength, short packet communication, URLLC, cooperative NOMA, max-min fairness.**

I. Introduction

The 5G systems and beyond, in addition to high throughput and capacity requests of traditional mobile broadband (MBB) services, should support new demands of achieving low latency and high reliability for many internet of things (IoT) applications. IoT applications are divided into two categories: massive machine-type communications (mMTC), and ultra-reliable low-latency communications (URLLC). The first one consists of many low-cost devices with massive connections and high battery lifetime requirements. While the second one's, URLLC, requirements are most related to mission-critical services in which uninterrupted and robust data exchange are vital.

To support low-latency communication, short packets with FBL codes are considered to reduce the transmission delay. In the FBL regime, in contrast to Shannon's capacity for infinite blocklength, decoding error probability at the receiver cannot be assumed negligible owing to short blocklength [1]. Polyanskiy et al. derived an exact approximation of information rate in the FBL regime at the AWGN channel [2]. Following that, research in this context developed to Multiple input multiple output (MIMO) channel with quasi-static fading [3] and quasi-static fading channel with retransmissions [4]. In [5], optimal power and blocklength allocation was considered in a high SNR scenario and the amount of NOMA transmission delay reduction was determined compared to orthogonal multiple access (OMA) in a closed-form. In [6], transmission rate and power allocation of the NOMA scheme were optimized to maximize the effective throughput of the strong user while the throughput of the other user was guaranteed at a certain level. A hybrid transmission scheme that combines time division multiple access (TDMA) and NOMA was proposed in [7], where the energy of the transmitter was minimized subject to heterogeneous latency constraints at receivers. In [8], to achieve max-min throughput in a two-user DL-NOMA system, an optimal power allocation algorithm was proposed.

In [9], relaying performance in the FBL regime was studied and the overall error probability of relaying along with its advantages over the direct transmission was investigated. The throughput and effective capacity of a relaying system in the FBL regime were obtained in [10] at the presence of a quasi-static fading channel and average channel state information (CSI) at the transmitter. In [11], a multi-terminal URLLC network was considered and the network reliability with multi-hop cooperative diversity was investigated in the FBL regime. A multi-relay system with the best relay selection approach was proposed in [12] for the FBL regime, and the achievable throughput bound was calculated using the polar codes. In [13], the author considered the cooperative relaying scenario with perfect CSI for a quasi-static Rayleigh fading channel and derived the outage probability of two-hop DF, SC, and MRC protocols. Optimization problems of average throughput and max-min throughput were studied in [14] using power and blocklength allocation under delay and consumed energy constraints by full search method with high complexity, but users' reliability was not guaranteed. Ren et al. in [15], considered optimal power and blocklength allocation in various transmissions schemes such as OMA, NOMA, relaying, and C-NOMA, to minimize the decoding error probability of the weak user, meanwhile, the reliability of the strong user's performance was guaranteed at a certain level.

In this work, we consider a DL transmission with two NOMA users and apply the cooperative relaying technique in the short packet communications scenario. The strong user, which performs successive interference cancellation (SIC) and detects data of the weak user, acts as a relay. The weak user receives its data via both BS and relay and implements SC protocol. Moreover, to guarantee the quality of service (QoS) of the weak user and to improve fairness, joint power and blocklength optimization is done to maximize the minimum throughput of the two users under latency, reliability, and energy constraints. Finally, to figure out advantages of the C-NOMA scheme in the FBL regime, we compare it with the optimal NOMA scenario proposed in [8].

II. Preliminaries Issues

A. System Model

As shown in Fig. 1(a), we consider a cooperative relaying scenario in a DL system with one BS and two NOMA users. In phase I, i.e., NOMA phase, BS transmits a NOMA frame of length m^{I} symbols, which consists of two users' data (N_1 bits, user 1's data and N_2 bits, user 2's data). User 1, the strong user, performs the SIC technique and decodes user 2's data and sends that to user 2 in a frame of length m^{II} symbols in phase II, i.e., relaying phase. The instantaneous channel coefficients of BS-user 1, BS-user 2, and user 1-user 2 links are denoted as h_1, h_2, and h_{12}, respectively. It is assumed that the channels are quasi-static Rayleigh fading. Hence, they are constant during one frame and vary independently from one frame to the next one. According to the power domain NOMA principle, in a two-user scenario, BS transmits $\sum_{i=1}^{2}\sqrt{p_i^{\mathrm{I}}}x_i$, where x_i is the message of user i, $i\in\{1,2\}$, and p_i^{I} refers to the allocated power of user i in phase I. So, the received signal at user i is given by $y_i^{\mathrm{I}}=(\sqrt{p_1^{\mathrm{I}}}x_1+\sqrt{p_2^{\mathrm{I}}}x_2)h_i+n_i$, where n_i is the complex additive white Gaussian noise with variance σ^2. Without loss of generality, it is assumed that $|h_1|^2>|h_2|^2$, and more power should be allocated to user 2. Therefore, user 1 can perform the SIC technique to remove the interference, while user 2 suffers from the interference and cannot cancel it. If x_2 is decoded correctly by user 1, it is re-encoded and transmitted ($\sqrt{p_2^{\mathrm{II}}}x_2'$ signal).[1] Consequently, the received signal at user 2 in the relaying phase is $y_2^{\mathrm{II}}=\sqrt{p_2^{\mathrm{II}}}x_2'\ h_{12}+n_{12}$. Let p_2^{II} show the allocated power to user 2 by the relay (user 1) in phase II, and n_{12} is the complex additive white Gaussian noise with variance σ^2 . To implement this scheme, like the one in [15], user 1 must know whether SIC is successful or not.

B. Direct Transmission Analysis in The FBL Regime

According to [2], the achievable data rate R for a finite blocklength of m symbols ($m\geq 100$), and an acceptable block error rate (BLER) ε , has an exact approximation as

$$R\approx C-\sqrt{\frac{V}{m}}\frac{Q^{-1}(\varepsilon)}{\ln 2} \tag{1}$$

where $C=\log_2(1+\gamma)$ is the Shannon capacity, γ is the SNR/SINR ratio, $Q^{-1}(\cdot)$ refers to the inverse Gaussian Q-function $Q(x)=\frac{1}{\sqrt{2\pi}}\int_x^{\infty}e^{-\frac{t^2}{2}}\,dt$, and $V=1-(1+\gamma)^{-2}$ is the channel dispersion. In the FBL regime, even with perfect CSI, the transmission is not error-free and the decoding error probability is given by

$$\varepsilon\approx Q\left(\frac{(C-R)\ln 2}{\sqrt{V/m}}\right)\triangleq Q(f(\gamma,R,m)) \tag{2}$$

[1]One should notice that x_2 is user 2's data with rate N_2/m^{I}, while x_2' is the same data with rate N_2/m^{II}.

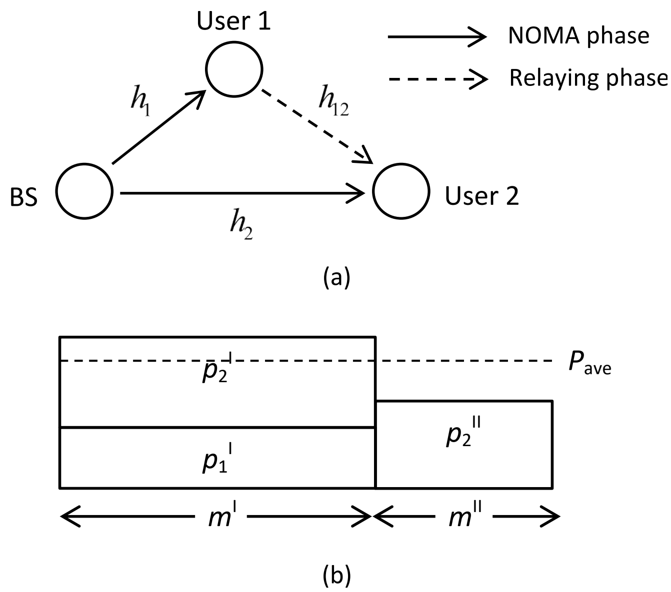

Fig. 1: (a) system model, (b) frame structure.

III. Performance Analysis of C-NOMA Transmission

It is assumed that the receivers have access to perfect CSI, and BS and each of the users have one antenna. Also, user 2 can employ SC strategy. In phase I, user 2 directly detects x_2 by considering x_1 as interference. The decoding error probability of x_2 at user 2 in phase I is denoted by $\varepsilon_{2,2}^{\mathrm{I}}$, which is approximated based on (2) by

$$\varepsilon_{2,2}^{\mathrm{I}}\approx Q(f(\gamma_{2,2}^{\mathrm{I}},R_{2,2}^{\mathrm{I}},m^{\mathrm{I}})) \tag{3}$$

where $\gamma_{2,2}^{\mathrm{I}}=p_2^{\mathrm{I}}|h_2|^2/(p_1^{\mathrm{I}}|h_2|^2+\sigma^2)$ and $R_{2,2}^{\mathrm{I}}=N_2/m^{\mathrm{I}}$ are the received SINR and the achievable rate of user 2 related to detecting x_2 in phase I, respectively. Since x_2 is detected directly, $\varepsilon_{2,2}^{\mathrm{I}}$ is the overall error probability of user 2 in phase I, i.e., $\varepsilon_2^{\mathrm{I}}=\varepsilon_{2,2}^{\mathrm{I}}$. On the opposite, user 1 performs SIC, meaning it first decodes x_2 while treats x_1 as interference. Similarly, the decoding error probability of x_2 at user 1 in phase I, which is denoted by $\varepsilon_{1,2}^{\mathrm{I}}$, is approximated as

$$\varepsilon_{1,2}^{\mathrm{I}}\approx Q(f(\gamma_{1,2}^{\mathrm{I}},R_{1,2}^{\mathrm{I}},m^{\mathrm{I}})) \tag{4}$$

where $\gamma_{1,2}^{\mathrm{I}}=p_2^{\mathrm{I}}|h_1|^2/(p_1^{\mathrm{I}}|h_1|^2+\sigma^2)$ and $R_{1,2}^{\mathrm{I}}=N_2/m^{\mathrm{I}}$ are the received SINR and the achievable rate of user 1 related to detecting x_2 in phase I, respectively. If user 1 decodes and removes x_2 successfully, then x_1 can be detected without interference. Accordingly, the decoding error probability of x_1 at user 1 in phase I, i.e., $\varepsilon_{1,1}^{\mathrm{I}}$, is denoted by

$$\varepsilon_{1,1}^{\mathrm{I}}\approx Q(f(\gamma_{1,1}^{\mathrm{I}},R_{1,1}^{\mathrm{I}},m^{\mathrm{I}})) \tag{5}$$

where $\gamma_{1,1}^{\mathrm{I}}=p_1^{\mathrm{I}}|h_1|^2/\sigma^2$ and $R_{1,1}^{\mathrm{I}}=N_1/m^{\mathrm{I}}$ are the received SINR and the achievable rate of user 1 related to detecting x_1 in phase I, respectively. By assuming that x_1 is detected when SIC is successful and the fact that in URLLC services, ε is usually in order of $10^{-5} \sim 10^{-9}$, the overall decoding error probability at user 1 in phase I can be approximated as

$$\varepsilon_1^{\mathrm{I}} = \varepsilon_{1,2}^{\mathrm{I}} + (1-\varepsilon_{1,2}^{\mathrm{I}})\varepsilon_{1,1}^{\mathrm{I}} \approx \varepsilon_{1,2}^{\mathrm{I}} + \varepsilon_{1,1}^{\mathrm{I}}. \tag{6}$$

Hence, the overall decoding error probability at user 1 is denoted as $\varepsilon_1 = \varepsilon_1^{\mathrm{I}}$.

In contrast, the overall decoding error probability of user 2 depends on the combining strategy. In SC protocol, user 2 does not combine the NOMA phase and relaying phase signals, but decodes transmitted messages from BS and relay (user 1) separately and selects the correctly decoded packet. First, the received message from user 1 in the relaying phase is decoded. If decoding is failed or no signal is received from user 1, then the transmitted message from BS in the NOMA phase is decoded. Therefore, an error occurs when both transmissions are unsuccessful. Decoding error probability of x_2' by user 2 in phase II, i.e., $\varepsilon_{2,2}^{\mathrm{II}}$, is given by

$$\varepsilon_{2,2}^{\mathrm{II}} \approx Q(f(\gamma_{2,2}^{\mathrm{II}}, R_{2,2}^{\mathrm{II}}, m^{\mathrm{II}})) \tag{7}$$

where $\gamma_{2,2}^{\mathrm{II}} = p_2^{\mathrm{II}}|h_{12}|^2/\sigma^2$ and $R_{2,2}^{\mathrm{II}} = N_2/m^{\mathrm{II}}$ are the received SNR and the achievable rate of user 2 related to detecting x_2' in phase II, respectively. One should note that the phase II signal will be transmitted if the message of user 2 is decoded correctly in phase I, so the overall decoding error probability of user 2 in phase II is approximated as

$$\varepsilon_2^{\mathrm{II}} = \varepsilon_{1,2}^{\mathrm{I}} + (1-\varepsilon_{1,2}^{\mathrm{I}})\varepsilon_{2,2}^{\mathrm{II}} \approx \varepsilon_{1,2}^{\mathrm{I}} + \varepsilon_{2,2}^{\mathrm{II}}. \tag{8}$$

Finally, the overall decoding error probability of user 2 in SC strategy is formulated as

$$\varepsilon_2 = \varepsilon_2^{\mathrm{I}}\varepsilon_2^{\mathrm{II}} \approx \varepsilon_{2,2}^{\mathrm{I}}(\varepsilon_{1,2}^{\mathrm{I}} + \varepsilon_{2,2}^{\mathrm{II}}). \tag{9}$$

IV. Problem Formulation

In the considered URLLC system, the two users are served with the aim of fairness during two phases with a total $D_{\max}$ symbols period. The throughput of user i, T_i, is defined as the average bits per each channel use (or complex symbol), which is decoded correctly at the receiver;

$$T_i \triangleq \frac{m^{\mathrm{I}}}{D_{\max}} R_{i,i}^{\mathrm{I}}(1-\varepsilon_i) \tag{10}$$

where $1-\varepsilon_i$ is the reliability of user i and a predefined value for each URLLC use case.

In the C-NOMA scheme, the superposition coding is performed in the NOMA phase, such that the BS enables to transmit users' signals simultaneously with different powers within a frame of length m^{I}. User 1 after decoding user 2's data, sends it in the relaying phase within a frame of length m^{II}. In Fig. 1(b) the frame structure of C-NOMA is observed. Therefore, the desired optimization problem is formulated as

$$\max_{\{p_i^{\mathrm{I}}, p_2^{\mathrm{II}}, m^j\}_{\substack{i=1,2,\\ j=\mathrm{I},\mathrm{II}}}} \quad \min\{T_1, T_2\} \tag{11a}$$

$$\text{s.t.} \quad m^{\mathrm{I}}\left(p_1^{\mathrm{I}} + p_2^{\mathrm{I}}\right) + m^{\mathrm{II}} p_2^{\mathrm{II}} \leq D_{\max} P_{\mathrm{ave}}, \tag{11b}$$

$$0 < p_1^{\mathrm{I}} + p_2^{\mathrm{I}} \leq \kappa_{\mathrm{p}} P_{\mathrm{ave}}, \ p_i^{\mathrm{I}} > 0, \ i \in \{1,2\}, \tag{11c}$$

$$0 \leq p_2^{\mathrm{II}} \leq \kappa_{\mathrm{p}} P_{\mathrm{ave}}, \tag{11d}$$

$$\varepsilon_i \leq {\varepsilon_i}^{\mathrm{th}}, \ i \in \{1,2\}, \tag{11e}$$

$$m^{\mathrm{I}} + m^{\mathrm{II}} = D_{\max}. \tag{11f}$$

Optimization parameters consist of blocklength and power allocated to two users in phases I and II. Constraint (11b) indicates the system's total energy consumption budget. Constraints (11c) and (11d) are the general power constraints, where P_{ave} is the average power, and κ_{p} is the peak to average power ratio (PAPR) factor. Constraint (11e) guarantees that the decoding error probability of user i does not violate $\varepsilon_i^{\mathrm{th}}$. Moreover, the latency constraint is stated by (11f).

V. Problem Solving

This section will solve the optimization problem in (11). To facilitate this issue, we first have to analyze the constraints and specify their optimal status. Let us first consider the constraint (11e) on the acceptable BLER of the two users. Since each URLLC use case needs specific reliability, allocating more resources to achieve a BLER lower than the required $\varepsilon_i^{\mathrm{th}}$, wastes the rare resources. Moreover, according to (1), a lower desired error probability results in a lower data rate. Therefore, $\varepsilon_i = \varepsilon_i^{\mathrm{th}}$ is an optimal choice. About constraint (11b), invoking [8, proposition 1], the acceptable data rate (i.e., $R > 0$) in (1), is a monotonically increasing function of the corresponding SNR/SINR. Using the contradiction method, one can prove that to maximize the throughput, the energy constraint holds with equality [15], i.e., $m^{\mathrm{I}}\left(p_1^{\mathrm{I}} + p_2^{\mathrm{I}}\right) + m^{\mathrm{II}} p_2^{\mathrm{II}} = D_{\max} P_{\mathrm{ave}}$. In addition, the following proposition indicates the ratio of optimal consumed energy in two transmission phases.

Proposition 1: At the optimal solution, the total consumed energy of the two users in phase I is always greater than the consumed energy in phase II, i.e., $m^{\mathrm{I}} P_{\mathrm{sum}} > m^{\mathrm{II}} p_2^{\mathrm{II}}$, where $P_{\mathrm{sum}} \triangleq p_1^{\mathrm{I}} + p_2^{\mathrm{I}}$. (The proof is eliminated due to the page limit.)

Furthermore, invoking [8, proposition 2], at the optimum point of Problem (11), throughputs of the two users are equal, i.e., $T_1 = T_2$. Following the above discussion, we provide a solution for the optimization problem in (11).

A. Optimal Design of Max-Min Fairness in C-NOMA

Since at the optimal solution $T_1 = T_2$, equation $R_{2,2}^{\mathrm{I}} = \frac{1-\varepsilon_1^{\mathrm{th}}}{1-\varepsilon_2^{\mathrm{th}}} R_{1,1}^{\mathrm{I}}$ can be derived from (10). Moreover, the message of user 2 contains the same number of bits in both phases, so it can be concluded that $R^{\mathrm{II}}_{2,2} = \frac{m^{\mathrm{I}}}{m^{\mathrm{II}}} R^{\mathrm{I}}_{2,2}$. Consequently, the optimization problem in (11) is rewritten as follows

$$\max_{\{m^{\mathrm{I}},p_1^{\mathrm{I}},p_2^{\mathrm{II}}\}} T_1 = \frac{m^{\mathrm{I}}}{D_{\max}} \left(1-\varepsilon_1^{\mathrm{th}}\right) R^{\mathrm{I}}_{1,1} \tag{12a}$$

$$\text{s.t.} \quad m^{\mathrm{I}} P_{\mathrm{sum}} + m^{\mathrm{II}} p_2^{\mathrm{II}} = D_{\max} P_{\mathrm{ave}} \tag{12b}$$

$$0 < P_{\mathrm{sum}} \le \kappa_{\mathrm{p}} P_{\mathrm{ave}},\ 0 < p_1^{\mathrm{I}} < \tfrac{P_{\mathrm{sum}}}{2} \tag{12c}$$

$$0 \le p_2^{\mathrm{II}} \le \kappa_{\mathrm{p}} P_{\mathrm{ave}},\ m^{\mathrm{I}} P_{\mathrm{sum}} > m^{\mathrm{II}} p_2^{\mathrm{II}} \tag{12d}$$

$$\varepsilon_i = {\varepsilon_i}^{\mathrm{th}},\ i \in \{1,2\} \tag{12e}$$

$$m^{\mathrm{I}} + m^{\mathrm{II}} = D_{\max}. \tag{12f}$$

The restriction on p_1^{I} in (12c) is applied based on the assumption that $|h_1|^2 > |h_2|^2$. So, to perform SIC correctly in the NOMA phase, it is necessary that $p_2^{\mathrm{I}} > p_1^{\mathrm{I}}$. This problem can be solved using exhaustive linear search; however, we shorten more the search range of to reduce the computational complexity. The main idea can be summarized as follows:

- First, by considering user 1's decoding error probability, i.e., $\varepsilon_1 \approx \varepsilon^{\mathrm{I}}_{1,2} + \varepsilon^{\mathrm{I}}_{1,1}$, the p_1^{I} bound that guarantees $\varepsilon_1 \le {\varepsilon_1}^{\mathrm{th}}$ is determined. According to our previous work in [8], ε_1 is convex in p_1^{I} and at most two values hold the $\varepsilon_1(p_1^{\mathrm{I}}) = {\varepsilon_1}^{\mathrm{th}}$. With $R^{\mathrm{I}}_{1,1} = 0$ and constant values of m^{I} and P_{sum}, we obtain the possible solutions that keep this equality in the range of $0 < p_1^{\mathrm{I}} < \frac{P_{\mathrm{sum}}}{2}$. Clearly, $\varepsilon^{\mathrm{I}}_{1,1}$ is a monotonically decreasing function of p_1^{I}, so it is derived that $p_1^{\mathrm{I,min}} = \arg\{\varepsilon_1(p_1^{\mathrm{I}}) \approx \varepsilon^{\mathrm{I}}_{1,1}(p_1^{\mathrm{I}}) = \varepsilon_1^{\mathrm{th}}\}$. On the other hand, $\varepsilon^{\mathrm{I}}_{1,2}$ a monotonically increasing function of p_1^{I} yields to $p_1^{\mathrm{I,max}} = \arg\{\varepsilon_1(p_1^{\mathrm{I}}) \approx \varepsilon^{\mathrm{I}}_{1,2}(p_1^{\mathrm{I}}) = \varepsilon_1^{\mathrm{th}}\}$. Hence, the search region of p_1^{I} is given by $p_1^{\mathrm{I,min}} \le p_1^{\mathrm{I}} \le p_1^{\mathrm{I,max}}$.
- Since the decoding error probability is a monotonically increasing function of the transmission rate, for each value of p_1^{I} in the feasible range, $R^{\mathrm{I}}_{1,1}$ is increased until user 1's decoding error probability equals to $\varepsilon_1^{\mathrm{th}}$. One should note that $R^{\mathrm{I}}_{1,1} \le C(\gamma^{\mathrm{I}}_{1,1})$.
- Only those $p_1^{\mathrm{I,min}} \le p_1^{\mathrm{I}} \le p_1^{\mathrm{I,max}}$ that satisfy $\varepsilon_2(p_1^{\mathrm{I}}) = {\varepsilon_2}^{\mathrm{th}}$ could be acceptable. Since the decoding error probability of user 2 in (9) is increasing function of p_1^{I}, the transmit power can be obtained using the bisection search method.
- After the full search on the values of m^{I} and P_{sum}, among the feasible solutions, the answer that maximizes T_1 is optimal.

Based on the above analysis, the algorithm for solving Problem (12) is proposed in Algorithm 1. It first determines the local maximum of T_1, i.e., ${T_0}^{\dagger}$, by taking constant m^{I} and checking all possible values of P_{sum} and p_1^{I}. In each iteration, the bisection search is adopted to find the desired p_1^{I}. By repeating this process on all possible m^{I} with a positive integer value, the global maximum of T_1, i.e., ${T_0}^{*}$, is found. Thus, using a three-dimensional (3-D) exhaustive linear search, the optimal global solution is achieved.

B. Computational Complexity

The computational complexity of Algorithm 1 is calculated as follows. In the first step, to obtain the bounds of p_1^{I}, a linear search with complexity Ω_1 is applied. In the next step, $R^{\mathrm{I}}_{1,1}$ is derived via the bisection method with complexity around $\log_2(\varepsilon_1^{\mathrm{th}}/\epsilon)$ where ϵ is the desired accuracy. Besides, the complexity of computing ε_2 is denoted as Ω_2. This step is performed at most $K_1 = (p_1^{\mathrm{I,max}} - p_1^{\mathrm{I,min}})/\Delta p$ times where Δp is the search step, so its complexity is denoted as $K_1\left(\log_2(\varepsilon_1^{\mathrm{th}}/\epsilon) + \Omega_2\right)$. In the last step, finding p_1^{I} via the bisection search method has complexity around $\log_2(\varepsilon_2^{\mathrm{th}}/\epsilon)$. These three steps are repeated on the possible values of P_{sum} and m^{I}, respectively $K_2 = \kappa_{\mathrm{p}} P_{\mathrm{ave}}/\Delta p$ and $D_{\max}$ times. Therefore, the worst-case complexity of Algorithm 1 is $\mathcal{O}\left(K_2 D_{\max}\left(\Omega_1 + K_1(\log_2(\varepsilon_1^{\mathrm{th}}/\epsilon) + \Omega_2) + \log_2(\varepsilon_2^{\mathrm{th}}/\epsilon)\right)\right)$.

Algorithm 1: Optimum Power and Blocklength Allocation Algorithm in the C-NOMA Scheme with SC Strategy.

Input: total blocklength $D_{\max}$, overall BLER of user i ${\varepsilon_i}^{\mathrm{th}}$, BS average power P_{ave}, required accuracy ϵ.
Output: optimum power $p_1^{\mathrm{I}*}$, $p_2^{\mathrm{I}*}$, $p_2^{\mathrm{II}*}$, and blocklength $m^{\mathrm{I}*}$, $m^{\mathrm{II}*}$, and fair throughput $T_1 = T_2 = {T_0}^{*}$.

```
for m^I = 1 : D_max do
    for P_sum = 0 : Δp : κ_p P_ave do
        -Set m^II := D_max − m^I and
         p_2^II := (D_max P_ave − m^I P_sum)/m^II.
        if 0 ≤ p_2^II ≤ κ_p P_ave & m^I P_sum ≥ m^II p_2^II then
            -Calculate p_1^{I,min} and p_1^{I,max}.
            -Set p_1^I := p_1^{I,min}.
            while ε_2 < ε_2^th do
                -Set p_1^I := min(p_1^I + Δp, p_1^{I,max}).
                -Find R_{1,1}^{I†} = arg{ε_1 = ε_1^th} via
                 bisection method with accuracy ϵ.
                -Calculate ε_2 by (9).
            -Set p_1^{I,lb} := p_1^I − Δp and p_1^{I,ub} := p_1^I.
            -Find p_1^{I†} ∈ [p_1^{I,lb}, p_1^{I,ub}] that satisfies
             ε_2 = ε_2^th via bisection method with
             accuracy ϵ.
    -Set R_{1,1}^{I‡} := max{R_{1,1}^{I†} | ε_2 = ε_2^th} and
     T_0^† := (1 − ε_1^th) m^I R_{1,1}^{I‡}/D_max.
-Set T_0^* := max{T_0^†}.
-Return: {m^{I*}, p_1^{I*}, p_2^{II*}} = arg max{T_0^†},
  m^{II*} = D_max − m^{I*}, p_2^{I*} = (D_max P_ave − m^{II*} p_2^{II*})/m^{I*} − p_1^{I*}.
```

VI. Numerical Results

In this section, the proposed C-NOMA scheme's performance with SC strategy is evaluated through the numerical results based on our analytical solutions. A heterogeneous network consists of URLLC users with different reliability requirements is considered. PAPR factor and required accuracy

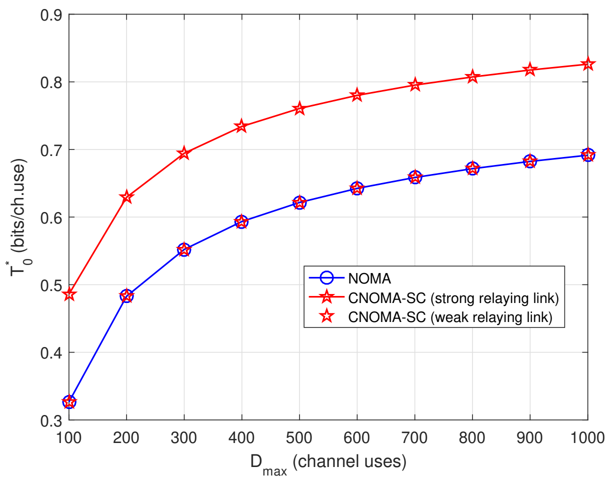

Fig. 2: Maximum fair throughput achieved by the C-NOMA and NOMA schemes versus $D_{\max}$, when $P_{\text{ave}} = 10$ W.

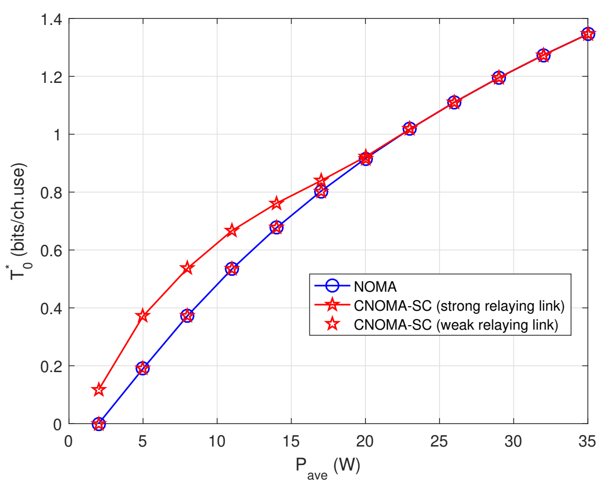

Fig. 3: Maximum fair throughput achieved by the C-NOMA and NOMA schemes versus P_{ave}, when $D_{\max} = 200$.

in Algorithm 1 are considered as $\kappa_{\text{p}} = 1.2$ and $\epsilon = 10^{-15}$, respectively. Also, it is assumed that $P_{\text{ave}} = 10$ W and $D_{\max} = 200$ channel uses, unless otherwise stated.

Moreover, the normalized channel gains of the NOMA phase and relaying phase are set to be fixed. For instance, it is assumed that $|h_1|^2/\sigma^2 = 0.8$ and $|h_2|^2/\sigma^2 = 0.1$. We investigate the performance of the proposed scheme in two status, i.e., strong relaying link with $|h_{12}|^2/\sigma^2 = 0.5$ and weak relaying link with $|h_{12}|^2/\sigma^2 = 0.01$.

In Fig. 2, the effect of total blocklength, $D_{\max}$, on the fair throughput in the proposed C-NOMA with SC receiver is assessed in two relaying link modes. Also, the results of optimal NOMA proposed in our previous work [8] are shown for comparison. It is observed that in the strong relaying link condition, the proposed C-NOMA effectively improves the fair throughput compared to the NOMA regardless of the blocklength. On the other hand, in a weak relaying link, the C-NOMA scheme has exactly the same performance as the NOMA. In fact, in this case, the optimal decision is in favor of the direct link, and the C-NOMA is transformed into the NOMA. However, in a realistic wireless channel, mixed conditions take place at the same time, and C-NOMA outperforms the NOMA on average.

In Fig. 3, the effect of average total power, P_{ave}, on the fair throughput is investigated. In the strong relaying link mode, the C-NOMA with SC strategy outperforms the NOMA in low power/SNR ranges, while it coincides with the NOMA on average powers greater than 20 W. This could be justified by the fact that in SC strategy, the signals don't combine, and transmission in phase II assures the success of user 2's packet decoding. Hence, in low SNRs where the weak user's probability of successful decoding in phase I is not too high, the reliability is increased by retransmission in phase II. However, in high SNRs where the allocated power of user 2 in NOMA phase guarantees the reliability, phase II transmission is pointless. Thus, in this case, transmission via a single phase is optimal in comparison with two-phase, and the proposed scheme performs like the NOMA. Moreover, in the weak relaying link mode, the C-NOMA scheme always complies with the NOMA. As a result, from the complexity perspective, the usage of C-NOMA with SC strategy seems sensible just in low SNR regimes.

VII. Conclusion

In this paper, the combination of NOMA with cooperative relaying technique (i.e., C-NOMA) was considered in short packet communications to guarantee high reliability and low latency particularly in low SNR scenarios. The performance of the SC strategy was presented in terms of decoding error probability in a quasi-static channel. Besides, the necessity to provide QoS of all users with critical services motived us to consider the max-min fairness as a design criterion. To this end, an optimization problem was formulated for a two-user DL C-NOMA system, and optimal power, blocklength, and transmission rate were determined under the total energy consumption, reliability, and delay constraints. Numerical results showed that the proposed C-NOMA improves the users' fair throughput compared to the NOMA in low SNRs and provides the same performance in higher SNRs. However, some concepts like extending the issue to other combining techniques and jointly design of users clustering and transmission strategies in multi-user environments remained for our future works.

References

[1] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, 2016.

[2] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[3] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4232–4265, 2014.

[4] P. Wu and N. Jindal, “Coding versus ARQ in fading channels: How reliable should the phy be?” *IEEE Transactions on Communications*, vol. 59, no. 12, pp. 3363–3374, 2011.
[5] Y. Yu, H. Chen, Y. Li, Z. Ding, and B. Vucetic, “On the performance of non-orthogonal multiple access in short-packet communications,” *IEEE Communications Letters*, vol. 22, no. 3, pp. 590–593, 2018.
[6] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen, and Z. Zhong, “Short-packet downlink transmission with non-orthogonal multiple access,” *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4550–4564, 2018.
[7] Y. Xu, C. Shen, T. Chang, S. Lin, Y. Zhao, and G. Zhu, “Transmission energy minimization for heterogeneous low-latency noma downlink,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1054–1069, 2020.
[8] F. Salehi, N. Neda, and M.-H. Majidi, “Max-min fairness in downlink non-orthogonal multiple access with short packet communications,” *AEU - International Journal of Electronics and Communications*, vol. 114, p. 153028, 2020.
[9] Y. Hu, J. Gross, and A. Schmeink, “On the performance advantage of relaying under the finite blocklength regime,” *IEEE Communications Letters*, vol. 19, no. 5, pp. 779–782, 2015.
[10] Y. Hu, A. Schmeink, and J. Gross, “Blocklength-limited performance of relaying under quasi-static rayleigh channels,” *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4548–4558, 2016.
[11] Y. Hu, M. Serror, K. Wehrle, and J. Gross, “Finite blocklength performance of cooperative multi-terminal wireless industrial networks,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 5778–5792, 2018.
[12] Y. Hu, C. Schnelling, M. C. Gursoy, and A. Schmeink, “Multi-relay-assisted low-latency high-reliability communications with best single relay selection,” *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7630–7642, 2019.
[13] P. Nouri, H. Alves, and M. Latva-aho, “On the performance of ultra-reliable decode and forward relaying under the finite blocklength,” in *2017 European Conference on Networks and Communications (EuCNC)*, 2017, pp. 1–5.
[14] Y. Hu, M. C. Gursoy, and A. Schmeink, “Efficient transmission schemes for low-latency networks: Noma vs. relaying,” in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2017, pp. 1–6.
[15] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, “Joint power and blocklength optimization for urllc in a factory automation scenario,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 1786–1801, 2020.