Rate-Splitting Multiple Access With Finite Blocklength for Short-Packet and Low-Latency Downlink Communications

Yunnuo Xu, Yijie Mao, Member, IEEE, Onur Dizdar, Member, IEEE, and Bruno Clerckx, Fellow, IEEE

Abstract—Rate-Splitting Multiple Access (RSMA) is an emerging flexible and powerful multiple access for downlink multi-antenna networks. In this paper, we introduce the concept of RSMA into short-packet downlink communications. We design optimal linear precoders that maximize the sum rate with Finite Blocklength (FBL) constraints. The relations between the sum rate and blocklength of RSMA are investigated for a wide range of network loads and user deployments. Numerical results demonstrate that RSMA can achieve the same transmission rate as Non-Orthogonal Multiple Access (NOMA) and Space Division Multiple Access (SDMA) with smaller blocklengths (and therefore lower latency), especially in overloaded multi-antenna networks. Hence, we conclude that RSMA is a promising multiple access for low-latency communications.

Index Terms—FBL, NOMA, RSMA, SDMA, short-packet communications, transmission latency.

I. INTRODUCTION

5G brings new features and enhancements compared to the previous generations of wireless communication networks, including higher data rates, Ultra-Reliable Low-Latency Communication (URLLC) and Machine Type Communication (MTC) [1]. There are many latency-sensitive applications, such as industrial automation, smart grid and intelligent transportation. In order to reduce the transmission latency, short-packets with Finite Blocklength (FBL) codes are typically adopted [2]. This brings a stringent latency requirement to the physical layer.

In the pioneering work [3], Polyanskiy et al. provided information-theoretic limits on the achievable rate for given FBL and Block Error Rate (BLER). Inspired by [3], the influence of FBL on the performance of various point-to-point communication scenarios has been investigated in [4], [5]. The maximum channel coding rate for both quasi-static fading channels and Additive White Gaussian Noise (AWGN) channels was studied in [4] with constraints on BLER, blocklength and a long-term transmit power. In [5], the authors investigated the maximum achievable transmission rate over quasi-static Multiple-Input Multiple-Output (MIMO) fading channels under both perfect and imperfect Channel State Information at the Transmitter (CSIT)/Channel State Information at the Receiver (CSIR) settings.

Under the assumption of infinite blocklength, it is known that Non-Orthogonal Multiple Access (NOMA) utilizing Superposition Coding with Successive Interference Cancellation (SC-SIC) accommodates multiple users via non-orthogonal resource allocation [6]. NOMA is well-known to outperform Orthogonal Multiple Access (OMA) with infinite blocklength [7] and FBL [8] communications. In multi-user multi-antenna systems, however, multi-antenna NOMA is shown to be an inefficient strategy in terms of multiplexing gain and use of SIC receivers [9]. Recently, Rate-Splitting Multiple Access (RSMA) has emerged as a promising multiple access for downlink multi-antenna networks under infinite blocklength [10], since it unifies and outperforms Space Division Multiple Access (SDMA), OMA, NOMA and physical-layer multicasting in multi-antenna networks for a wide range of network loads and user deployments [11], [12]. Furthermore, RSMA provides significant gains in terms of spectral efficiency, energy efficiency, robustness and Channel State Information (CSI) feedback overhead reduction [12]–[14], which are also important in the context of low-latency communications.

Motivated by the benefits of RSMA in the infinite blocklength regime and the significant performance gains of RSMA in link-level evaluations with practical codes [15], we introduce RSMA into FBL downlink communications in this paper. All existing works of RSMA are based on the Shannon theorem with infinite blocklength. To the best of our knowledge, this is the first paper studying and optimizing RSMA for FBL communications. To that end, we leverage the information-theoretic limits provided in [3], and investigate the achievable sum rate under the constraints on blocklength and transmit power. Our results show that by utilizing RSMA for FBL downlink communications, the blocklength can be reduced, and the latency is therefore decreased, compared to SDMA and NOMA for the same transmission rate. Alternatively, for a given blocklength and therefore latency, RSMA can achieve a higher sum rate than SDMA and NOMA.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a system that consists of a Base Station (BS) equipped with \( N_t \) antennas and \( K \) single-antenna users. The users are indexed by \( \mathcal{K} = \{1, 2, \ldots, K\} \). The 1-layer RSMA scheme is adopted to serve \( K \) users in this paper. Assume that the BS intends to transmit \( K \) messages to \( K \) users. The message intended for user-\( k \) is denoted by \( W_k \), \( \forall k \in \mathcal{K} \). The \( W_k \) is divided into two parts, \( \{W_{c,k}, W_{p,k}\} \), which are named as the common part and the private part, respectively. Messages \( \{W_{c,1}, \ldots, W_{c,K}\} \) are encoded together into the common stream \( s_c \) to be decoded by all \( K \) users. The private message \( W_{p,k} \) is encoded into the private stream \( s_p \). The overall symbol stream vector to be transmitted is denoted by \( s = [s_c, s_1, s_2, \ldots, s_K]^T \). The symbols are precoded via a preceding matrix \( P = [p_c, p_1, p_2, \ldots, p_K] \), where \( p_j \in \mathbb{C}^{N_t \times 1} \) represents the linear precoder for the stream \( s_j \), \( j \in \{c, 1, 2, \ldots, K\} \). The transmit signal can be expressed by

\[
\mathbf{x} = \mathbf{P} \mathbf{s} = p_c s_c + \sum_{k \in \mathcal{K}} p_k s_k.
\]

(1)

We assume that the symbols have unit power, i.e., \( \mathbb{E} \{ |s|^2 \} = 1 \). The transmit signal is subject to a maximum transmit power \( P_t \), such that, \( \mathbb{E} \{ |\mathbf{x}|^2 \} \leq P_t \), and equivalently \( \text{tr} \{ \mathbf{P}^H \mathbf{P} \} \leq P_t \). The received signal \( y_k \) at user-\( k \) is written as

\[
y_k = \mathbf{h}_k^H \mathbf{x} + n_k,
\]

(2)

1-layer RSMA is denoted as RSMA in the rest of the paper unless otherwise stated. More complex Rate-Splitting (RS) architectures such as the generalized RS proposed in [12] can be considered for future works.

0018-9545 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
where $\mathbf{h}_k \in \mathbb{C}^{N \times 1}$ is the channel between the BS and user-$k$, and term $n_k \sim \mathcal{CN}(0, 1)$ is AWGN at user-$k$. We assume perfect CSIT and perfect CSIR.

Each user first decodes the common stream $s_c$ by treating the interference from the private streams $s_1, \ldots, s_K$ as noise. Accordingly, the Signal to Interference plus Noise Ratio (SINR) of the common stream at user-$k$ is

$$
\Gamma_{c,k} = \frac{|\mathbf{h}_{ck}^H \mathbf{p}_c|^2}{\sum_{j \in K \setminus \{c\}} |\mathbf{h}_{jk}^H \mathbf{p}_j|^2 + 1}.
$$

Assuming the common stream $s_c$ is successfully decoded, it is reconstructed and removed from the received signal $y_k$. Subsequently, user-$k$ decodes the private stream $s_k$ while treating the private streams of other users as noise. Accordingly, the SINR of decoding the private stream $s_k$ at user-$k$ is

$$
\Gamma_{p,k} = \frac{|\mathbf{h}_{pk}^H \mathbf{p}_k|^2}{\sum_{j \in K \setminus \{p\}} |\mathbf{h}_{jk}^H \mathbf{p}_j|^2 + 1}.
$$

Once the common and private streams are decoded, user-$k$ reconstructs the original message by extracting the decoded $W_{c,k}$ from the decoded $W_c$, and combining the decoded $W_{c,k}$ with the decoded $W_{p,k}$. Following [3], the respective rates $R_{c,k}, R_{p,k}$ of the common and private streams are

$$
R_{c,k} \approx \log_2(1 + \Gamma_{c,k}) - (\log_2 e) \sqrt{\frac{V_{c,k}}{l_c}} Q^{-1}(\epsilon),
$$

$$
R_{p,k} \approx \log_2(1 + \Gamma_{p,k}) - (\log_2 e) \sqrt{\frac{V_{p,k}}{l_p}} Q^{-1}(\epsilon),
$$

where $l_c, l_1, \ldots, l_K$ are the respective blocklengths of streams $s_c, s_1, \ldots, s_K$ and $\epsilon$ represents the BLER. The function $Q^{-1}(\cdot)$ corresponds to the inverse of the Gaussian Q function. $V_{c,k}$ and $V_{p,k}$ are the channel dispersion parameters with expressions given by

$$
V_{i,k} = 1 - (1 + \Gamma_{i,k})^{-2}, \ i \in \{c, p\}.
$$

In this work, we consider signaling by Gaussian codebooks, so that the expression (5) is valid for our setup. To ensure that the common stream is successfully decoded by all users, the achievable rate of the common stream (also denoted as common rate) shall not exceed $R_c = \min\{R_{c,1}, R_{c,2}, \ldots, R_{c,K}\}$.

We focus on the precoder design for FBL signal $x$ to maximize the sum rate subject to blocklength and transmit power constraints. The blocklength for each stream is $l_i$, i.e., $l_i = l$, $j \in \{c, 1, 2, \ldots, K\}$. The blocklength of the codeword is proportional to the latency and can be approximately expressed as, $l \approx BT$, where $B$ and $T$ are the bandwidth and time duration of the signal (i.e., latency), respectively [2]. In this work, we fix the bandwidth $B$ and investigate the effect of blocklength. For a given blocklength $l$, the achievable sum rate of $K$-user RSMA is

$$
\max_{P, c, \beta_p} \sum_{k \in K} (C_k + \beta_p) \quad \text{s.t.} \quad \sum_{k \in K} C_k' \leq R_{c,k}, \ \forall k \in K
$$

$$
\text{tr} \left( P P^H \right) \leq P_t
$$

$$
R_{c,k,tot} \geq r_{th}^c, \ \forall k \in K
$$

$$
c \geq 0.
$$

where $c = [C_1, C_2, \ldots, C_K]$ and $C_k$ is the $k$th user's portion of the common rate with $\sum_{k \in K} C_k = R_c$. The total transmission rate of user-$k$ is $R_{k,tot} = C_k + R_{p,k}$. The constraint (7d) guarantees the Quality-of-Service (QoS) for the users by the individual rate constraint $r_{th}^k, \ \forall k \in K$. The sum rate maximization problem (7) is a non-convex fractional program. Considering the asymptotic case when $l$ goes to infinity, the second terms in $R_{c,k}$ and $R_{p,k}$ become 0 and the optimization problem (7) is simplified to the one considered in [12] with infinite blocklength. Therefore, the optimization problem (7) is more general than the one investigated in [12]. By comparing sum rate expressions of FBL and infinite blocklength scenarios, it is observed that the sum rate loss results from second terms in the common and private rate.

### III. PROPOSED ALGORITHM

In this section, we describe the Successive Convex Approximation (SCA) algorithm adopted to jointly optimize precoders $P$ and the common rate vector $c$. SCA has been introduced in [16] for RSMA precoder design in cooperative systems. The proposed SCA approach differs from the existing one in [16] due to the FBL rate expressions in (5) and the non-convex nature of the channel dispersion expression, which brings further challenge to approximation.

The problem (7) is non-convex due to the non-convex rate expressions. We introduce variables $\beta_p = [\beta_{p,1}, \beta_{p,2}, \ldots, \beta_{p,K}]$, $\rho_c = [\rho_{c,1}, \rho_{c,2}, \ldots, \rho_{c,K}]$, $\rho_p = [\rho_{p,1}, \rho_{p,2}, \ldots, \rho_{p,K}]$, $\sigma_c = [\sigma_{c,1}, \sigma_{c,2}, \ldots, \sigma_{c,K}]$ and $\sigma_p = [\sigma_{p,1}, \sigma_{p,2}, \ldots, \sigma_{p,K}]$, where $\beta_{p,k}$ is the lower bound of the private rate $R_{p,k}$, $\rho_{c,k}$ and $\rho_{p,k}$ are the lower bounds of the SINR of the common and private streams, respectively, and $\sigma_{c,k}$ and $\sigma_{p,k}$ are the upper bounds for the interference plus noise terms corresponding to $\rho_{c,k}$ and $\rho_{p,k}$, respectively. For a given BLER and blocklength, $D = \frac{Q^{-1}(\epsilon)}{l} \log_2 e$ is a constant. We also introduce the dispersion parameter in terms of $\rho_{i,k}$ as

$$
u_i = 1 - (1 + \rho_{i,k})^{-2}, i \in \{c, p\}.
$$

With the introduced variables, Problem (7) is equivalently written as (8).

$$
\max_{P, c, \beta_p, \rho_c, \sigma_c} \sum_{k \in K} (C_k + \beta_p) \quad \text{s.t.} \quad \log_2(1 + \rho_{c,k}) - D\sqrt{\tau_{c,k}} \geq \sum_{k' \in K} C_{k'}, \ \forall k \in K
$$

$$
\log_2(1 + \rho_{p,k}) - D\sqrt{\tau_{p,k}} \geq \beta_p, \ \forall k \in K
$$

$$
\frac{|\mathbf{h}_{ck}^H \mathbf{p}_c|^2}{\sigma_{c,k}} \geq \rho_{c,k}, \ \forall k \in K
$$

$$
\frac{|\mathbf{h}_{pk}^H \mathbf{p}_k|^2}{\sigma_{p,k}} \geq \rho_{p,k}, \ \forall k \in K
$$

$$
\sigma_{c,k} \geq \sum_{j \in K \setminus \{k\}} |\mathbf{h}_{jk}^H \mathbf{p}_j|^2 + 1, \ \forall k \in K
$$

$$
\sigma_{p,k} \geq \sum_{j \in K \setminus \{k\}} |\mathbf{h}_{jk}^H \mathbf{p}_j|^2 + 1, \ \forall k \in K
$$

$$
C_k + \beta_p \geq r_{th}^k, \ \forall k \in K
$$

$$
(7c), (7e).
$$

Problem (8) remains non-convex due to the non-convex constraints (8b)-(8e). Next, we approximate the non-convex parts $\sqrt{\tau_{c,k}}$ and $\sqrt{\tau_{p,k}}$ in the constraints by the first-order Taylor series. Constraints (8b)
Algorithm 1: SCA-Based Algorithm.

1. Initialize: $n \leftarrow 0$, $t^{[n]} \leftarrow 0$, $P^{[n]}$, $P_c^{[n]}$, $P_s^{[n]}$, $\sigma_c^{[n]}$, $\sigma_p^{[n]}$;
2. repeat
3. $n \leftarrow n + 1$;
4. Solve problem (11) using $P^{[n-1]}$, $P_c^{[n-1]}$, $P_s^{[n-1]}$, $\sigma_c^{[n-1]}$, $\sigma_p^{[n-1]}$, and denote the optimal value of the objective function as $t^n$ and the optimal solutions as $P_c^*$, $P_s^*$, $\sigma_c^*$, $\sigma_p^*$;
5. Update $t^{[n]} \leftarrow t^n$, $P^{[n]} \leftarrow P^*$, $P_c^{[n]} \leftarrow P_c^*$, $P_s^{[n]} \leftarrow P_s^*$, $\sigma_c^{[n]} \leftarrow \sigma_c^*$, $\sigma_p^{[n]} \leftarrow \sigma_p^*$;
6. until $|t^{[n]} - t^{[n-1]}| < \tau$;

IV. RESULTS AND DISCUSSION

The performances of RSMA, NOMA and SDMA with FBL in underloaded and overloaded scenarios are evaluated below.

A. Underloaded Scenario

1) Rate Comparison: In a two-user scenario, we first consider a simplified channel model of Uniform Linear Array (ULA) in order to investigate the effect of user angle and channel strength on the performance [12]. When $N_z = 4$, the channels between the BS and users are realized as $h_1 = [1, 1, 1, 1]^H$, $h_2 = \gamma \times [1, e^{\theta}, e^{2\theta}, e^{3\theta}]^H$, where $\gamma$ is the channel strength of user-2, and $\theta$ represents the difference in angle of departure between users.

2) SNR and Power Allocation: The performances of RSMA, NOMA and SDMA with FBL in underloaded and overloaded scenarios are evaluated below.

3) Performance Evaluation: The notations “Inf” and “Fin” in figures represent the schemes when $l = \infty$ and when $l$ is finite, respectively. In Fig. 1, the sum rate increases with the blocklength for all strategies. As the blocklength increases, the sum rate of all strategies approach that with infinite blocklength, as expected. As shown in Fig. 1, RSMA outperforms NOMA and SDMA in all cases, especially when $\theta = \frac{\pi}{6}$. For the channel angles $\theta \in \{\frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$, the performance of RSMA is almost identical with SDMA, and both achieve better sum rate performance than NOMA. Furthermore, when $\theta = \frac{\pi}{2}$, RSMA requires a blocklength of 100 to achieve a sum rate of 9.7 bit/s/Hz, while SDMA needs a significantly longer blocklength of around 2500 to achieve the same sum rate. The results show that RSMA brings great benefits in terms of blocklength reduction as the user channels get more aligned with each other.
We analyse the rate achieved by applying the precoder obtained under the infinite blocklength assumption to the scenario with FBL, i.e., “Inf-Fin RSMA”. From Fig. 2, the gain of “Fin RSMA” over “Inf-Fin RSMA” is observed especially when the blocklength is small, and it is more obvious with small SNR value.

2) Power Allocation: The relation between power allocation of RSMA and blocklength is investigated in Fig. 3(a) for $\theta \in \left\{ \frac{9}{4}, \frac{3}{2} \right\}$.

![Graph](image)

**Fig. 1.** Sum rate versus blocklength of different strategies for underloaded two-user deployment, $\gamma = 1$. (a) $\theta = \frac{9}{4}$, (b) $\theta = \frac{3}{2}$, (c) $\theta = \frac{5}{4}$, (d) $\theta = \frac{1}{2}$.

**Fig. 2.** Applying the precoder from infinite blocklength scenario to FBL scheme with different SNR in underloaded two-user deployment, $\theta = \frac{9}{4}$. (a) SNR = 0 dB, (b) SNR = 20 dB.

An important observation from Fig. 3(a) is that the power allocated to the common stream tends to increase with the blocklength. Such phenomenon is explained as follows. When FBL is considered, the second term $S(x) = (\log_2 e) \sqrt{\frac{T}{Q}} (\epsilon_{RSMA})$ (also known as the channel dispersion term in (5)) reduces the achievable rate. Compared to SDMA, the additional stream introduced by RS (i.e., common stream) calls for power re-allocation (i.e., a part of $P_1$ may be allocated to the common stream). The use of common stream results in an additional second term $S(V_{c,k})$ in the sum rate expression, as evident from (5a).

The splitting ratio is adjusted to compensate the term $S(V_{c,k})$ by decreasing the power allocated to the common stream (due to the monotonic decreasing behavior of the channel dispersion term with decreasing SINR). When the blocklength is reduced below a certain value, the power allocated to the common stream is adjusted and RS is not performed (i.e., zero power is allocated to the common stream), as observed in the case for $\theta = \frac{9}{4}$ with a blocklength of 500 bits. For a blocklength that is smaller than 500 bits, the sum rate of SDMA is higher than that obtained by forcing to perform RS. This is visualized in Fig. 3(b), which shows the term $S(V_c)$ for the common stream, where $V_c = \arg \min_{V_{c,k}} R_{c,k}$. We first consider the case with $\theta = \frac{3}{2}$, where RSMA achieves a sum rate gain of approximately 0.3 bit/s/Hz under infinite blocklength assumption, as observed from Fig. 1(a). When $l \geq 500$ and $\theta = \frac{3}{2}$, $S(V_c)$ increases as the blocklength decreases, while in the range of $l < 500$, RS is not performed due to the term $S(V_c)$ increasing beyond 0.3 bit/s/Hz. Consequently, the power allocated to the common stream is used for the private streams to achieve a better sum rate performance. The observation from Fig. 3(b) aligns with that from Fig. 3(a) and 1(a). On the other hand, the power allocated to the common stream does not change significantly with varying blocklength for $\theta = \frac{9}{4}$, as RSMA achieves a significant gain over SDMA and NOMA for the considered $\theta$ value with finite and infinite blocklength. When $\theta \geq \frac{3}{2}$, RSMA does not achieve any sum rate gain over SDMA with infinite blocklength coding, thus RS is not performed for FBL coding, either.

**B. Overloaded Scenario**

We investigate an overloaded scenario in this section. The performances of RSMA and SDMA are studied for special channel realizations and random channel realizations.

The BS is equipped with two antennas and serves four single-antenna users. The channels of users are realized as $H_1 = [1, 1]^H$, $h_2 = \gamma_1 \times [1, e^{j\theta_1}]^H$, $H_3 = [1, 1, 1, 1]^H$, $h_4 = \gamma_2 \times [1, e^{j\theta_2}]^H$, $h_5 = \gamma_3 \times [1, e^{j\theta_3}]^H$, $\gamma_1, \gamma_2, \gamma_3$ and $\theta_1, \theta_2, \theta_3$ are control variables. We assume that $\gamma_1 = \gamma_2 = \gamma_3 = 1$, $\theta_1 \in \left(0, \frac{\pi}{2}\right)$, $\theta_2 = \theta_1 + \frac{\pi}{2}$ and $\theta_3 = \theta_1 + \frac{\pi}{4}$ for Fig. 4. The individual rate constraint follows the same setting in Section IV-A.

Fig. 4 shows that “Fin RSMA” outperforms “Fin SDMA” when $\gamma_1 = 0.3$. Besides, “Fin RSMA” even outperforms “Inf SDMA” since
The channel
IEEE Trans. Inf. Theory
IEEE Trans. Wireless Commun.
0
ϕ
8. The channel
SDMA performance degrades significantly in overloaded scenarios. RSMA utilizes much smaller blocklength compared to SDMA for achieving the same sum rate. Hence, the latency is reduced greatly.

Fig. 5 shows sum rate performances of RSMA and SDMA in both infinite and finite blocklength scenarios under random channel realizations. We set \( N_t = 4 \), \( K = 8 \). The channel \( h_k \) has independent and identically distributed (i.i.d) complex Gaussian entries with a certain variance, i.e., \( CN(0, \sigma^2_k) \). \( \sigma^2_1 = 1, \sigma^2_2 = 0.875, \ldots, \sigma^2_8 = 0.125 \). The sum rate performances are obtained by averaging over 100 random channel realizations. The individual rate constraint is \( r_{th}^b = 0.2 \) bit/s/Hz for all blocklengths. The figure demonstrates that compared with finite SDMA, finite RSMA requires much smaller blocklength for achieving the same transmission rate. SDMA requires a blocklength of 2500 to achieve 20.4 bit/s/Hz, instead RSMA only needs a blocklength of 300 to achieve the same sum rate. RSMA can remarkably reduce blocklength (and therefore latency) in overloaded scenarios.

V. CONCLUSION

This paper introduces RSMA in FBL downlink communications with low latency requirements. An optimization problem is formulated to maximize the sum rate of users. A SCA-based algorithm is adopted to solve the optimization problem. The results show that in the FBL communications, RSMA can attain the same transmission rate with smaller blocklength (and hence lower latency) in comparison with NOMA and SDMA. Particularly, in overloaded multi-antenna networks, the latency of RSMA is significantly reduced. Consequently, we conclude that RSMA is a promising strategy for low-latency communications.

We use the assumption of perfect CSIT to shed light on how RSMA performs with FBL. There are extensive works studying RSMA performance with infinite blocklength and imperfect CSIT. The superiority of RSMA over conventional strategies was shown in terms of CSIT inaccuracy, degree of freedom and complexity. It is expected that RSMA with FBL would have even higher performance gain compared to other schemes. The performance of FBL RSMA with imperfect CSIT is an interesting problem worth future investigation.

REFERENCES

[1] P. Popovski, “Ultra-reliable communication in 5G wireless systems,” in Proc. IEEE 1st Int. Conf. 5G Ubiquitous Connectivity, 2014, pp. 146–151.
[2] G. Durisi, T. Koch, and P. Popovski, “Toward massive, ultra-reliable, and low-latency wireless communication with short packets,” IEEE Trans. Mobile Radio Commun., vol. 104, no. 9, pp. 1711–1726, Sep. 2016.
[3] Y. Polyanskiy, H. V. Poor, and S. Verdu, “Channel coding rate in the finite blocklength regime,” IEEE Trans. Inf. Theory, vol. 56, no. 5, pp. 2307–2359, May 2010.
[4] W. Yang, G. Caire, G. Durisi, and Y. Polyanskiy, “Optimum power control at finite blocklength,” IEEE Trans. Inf. Theory, vol. 61, no. 9, pp. 4598–4615, Sep. 2015.
[5] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, “Quasi-static multiple-antenna fading channels at finite blocklength,” IEEE Trans. Inf. Theory, vol. 60, no. 7, pp. 4232–4265, Jul. 2014.
[6] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chiti-Lin, and Z. Wang, “Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends,” IEEE Commun. Mag., vol. 53, no. 9, pp. 74–81, Sep. 2015.
[7] T. Cover, “Broadcast channels,” IEEE Trans. Inf. Theory, vol. 18, no. 1, pp. 2–14, Jan. 1972.
[8] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen, and Z. Zhong, “Short-packet downlink transmission with non-orthogonal multiple access,” IEEE Trans. Wireless Commun., vol. 17, no. 7, pp. 4550–4564, Jul. 2018.
[9] B. Clerckx et al., “Is NOMA efficient in multi-antenna networks? A critical look at next generation multiple access techniques,” IEEE Open J. Commun. Soc., vol. 2, pp. 1310–1343, 2021.
[10] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassouli, “Rate splitting for MIMO wireless networks: A promising PHY-layer strategy for LTE evolution,” IEEE Commun. Mag., vol. 54, no. 5, pp. 98–105, May 2016.
[11] B. Clerckx, Y. Mao, R. Schober, and H. V. Poor, “Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel: A simple two-user rate analysis,” IEEE Wireless Commun. Lett., vol. 9, no. 3, pp. 349–353, Mar. 2020.
[12] Y. Mao et al., “Rate-splitting multiple access for downlink communications: Bridging, generalizing, and outperforming SDMA and NOMA,” EURASIP J. Wireless Commun. Netw., vol. 2018, no. 1, pp. 1–54, 2018.
[13] C. Han, Y. Wu, and B. Clerckx, “Rate analysis of two-receiver MISO broadcast channel with finite rate feedback: A rate-splitting approach,” IEEE Trans. Commun., vol. 63, no. 9, pp. 3232–3246, Sep. 2015.
[14] H. Joudeh and B. Clerckx, “Sum-rate maximization for linearly precoded downlink multiuser MISO systems with partial CSIT: A rate-splitting approach,” IEEE Trans. Commun., vol. 64, no. 11, pp. 4847–4861, Nov. 2016.
[15] O. Ouzdar, Y. Mao, W. Han, and B. Clerckx, “Rate-splitting multiple access for downlink multi-antenna communications: Physical layer design and link-level simulations,” in Proc. IEEE 31st Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun., 2020, pp. 1–6.
[16] Y. Mao, B. Clerckx, J. Zhang, V. O. K. Li, and M. A. Arafah, “Max-min fairness of K-user cooperative rate-splitting in MISO broadcast channel with user relaying,” IEEE Trans. Wireless Commun., vol. 19, no. 10, pp. 6362–6376, Oct. 2020.
[17] Y. Mao and B. Clerckx, “Beyond dirty paper coding for multi-antenna broadcast channel with partial CSIT: A rate-splitting approach,” IEEE Trans. Commun., vol. 68, no. 11, pp. 6775–6791, Nov. 2020.

Fig. 5. Sum rate versus blocklength for different strategies in overloaded eight-user deployment.