Interference Cancellation via D2D CSI Sharing for MU-MISO-NOMA System with Limited Feedback

Mohanad M. Al-Wani, A. Sali, *Senior Member, IEEE*, Sumaya D. Awad, Asem A. Salah, Zhiguo Ding, *Fellow Member, IEEE*, Nor K. Noordin, S. J. Hashim, and Chee Yen Leow

***Abstract*—In downlink non-orthogonal multiple access (NOMA) with multiuser clustering, a single beamforming vector precodes the superimposed signal of strong and weak users. This precoding results in imperfect inter-cluster interference (ICI) cancellation at weak users since beamforming vectors are designed based on strong users' channels. Also, the ICI is highly increased with the number of transmit antennas, causing severe degradation in the performance of weak users and total system. In this paper, *first*, a new cooperative NOMA is introduced, in which weak user beam-matching (WBM) equalizer is proposed to remove the ICI at weak users by matching weak users' channels with the designed beams. The process of WBM is accomplished with the aid of device-to-device (D2D) channel state information (CSI) sharing between the nearby strong and weak users. *Second*, based on WBM principle, strong user beam-matching (SBM) equalizer is proposed to eliminate the generated ICI at strong users in the case of limited feedback. *Third*, a new power allocation strategy is proposed to improve weak users' performance by considering the gained throughput from interference cancellation. *Finally*, besides the sum-rate, which is adopted as the performance metric by most of the existing NOMA works, the bit error rate (BER) of NOMA users in cooperative NOMA is calculated and compared with those in other schemes. Simulation results show that our cooperative NOMA with the proposed equalizers achieves significant sum-rate and BER improvements over other non-cooperative schemes with both perfect and limited feedback scenarios.**

***Index Terms*—Cooperative NOMA, multiuser, beamforming, limited feedback, device-to-device (D2D), CSI sharing, BER.**

This work was supported in part by: NOMA-MIMO: Optimizing 5G Wireless Communication Performance Based on Hybrid NOMA with Partial Feedback for Multiuser MIMO (GP-IPS/2018/9663000, Vote No: 9663000), in part by: EMOSEN-Energy Efficient MIMO-Based Wireless Transmission for SWIPT-Enabled Network (Vot No.: 9671600) (UPM/800-3/3/1/GPB/2019/9671600), in part by 'ATOM' -Advancing the State of the Art of MIMO (Proj. No.: 690750-ATOM-H2020-MSCA-RISE-2015, UPM: 6388800-10801). (Corresponding authors: A. Sali; Mohanad M. Al-Wani)

Mohanad M. Al-Wani, A. Sali, Sumaya D. Awad, Nor K. Noordin and S. J. Hashim are with the Wireless and Photonic Networks Research Centre of Excellence (WiPNet), Department of Computer and Communication Systems Engineering, Faculty of Engineering, UPM, 43400 Serdang, Selangor (e-mails: moheng84anad@gmail.com; aduwati@upm.edu.my; sumayadhari@gmail.com; nknordin@upm.edu.my; sjh@upm.edu.my).

Asem A. Salah is with the Department of Computer System Engineering, Faculty of Engineering and Information Technology, Arab American University, Palestine, Jenin (e-mail: asem.salah@aauj.edu).

Zhiguo Ding is with the School of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, UK (e-mail: zhiguo.ding@manchester.ac.uk).

Chee Yen Leow is with Wireless Communication Centre, School of Electrical Engineering, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia (e-mail: bruceleow@utm.my).

I. Introduction

Future wireless networks such as beyond fifth-generation (B5G) or sixth-generation (6G) networks are expected to support extremely high data rates (up to 1 Tbps) with a diverse range of quality-of-service (QoS) requirements [1]–[3].

In order to offer significant improvements for cellular networks, the 6G wireless networks require an efficient multiple access technique. In the past few years, it has been shown that non-orthogonal multiple access (NOMA) can improve the spectral efficiency and support more users than the conventional orthogonal multiple access (OMA) schemes using the same wireless resources [4]. Therefore, NOMA has been recognized as one of the key technologies to meet the demands of the upcoming B5G wireless networks [5], [6].

The conventional OMA schemes are not able to support large wireless network capacity because orthogonal resources are allocated to different users. On the other hand, NOMA allows multiple users to share the same resources in terms of time, frequency and/or spatial domain through power-domain multiplexing and successive interference cancellation (SIC), which leads to better spectral efficiency and can support more users and/or devices than OMA schemes [4], [7], [8].

In downlink beamforming (BF) for multiuser multiple-input single-output (MU-MISO) system, the data rate and the number of supportable users can be further increased by integrating NOMA with this system (MU-MISO-NOMA), since in a conventional MU-MISO system, each BF vector is devoted to one user. Whereas in MU-MISO-NOMA system, a single BF vector can be shared by multiple users (clustered users). In [9], [10], NOMA was firstly combined with MU-MISO in a single cell network using zero-forcing beamforming (ZFBF) technique. In our previous works [11], [12], the user fairness was enhanced over those in [10] by proposing fair user clustering algorithms based on proportional fairness. In [11], the conventional user clustering algorithm in [10] is integrated with the proportional fairness selection criterion under the assumption of perfect channel state information at the transmitter (perfect CSIT). Whereas in [12], two fair user clustering algorithms were introduced, which can give better throughput-fairness-trade-off than that in [11], and the system was tested with both perfect CSIT and limited feedback cases.

In MU-MISO-NOMA [9]–[12], the BF vectors are designed based on the channels of strong users. Hence, the inter-cluster interference (ICI) can be completely removed at strong users. Whereas, weak users (cell-edge users) suffer from the residual ICI due to the mismatch between weak users' channels and the

designed BF vectors. This ICI strongly degrades the data rate and increases the error-rate at weak users, especially when the number of transmit antennas N_t increases. Therefore, those works have considered $N_t = 2$ only to have the minimum possible ICI at weak users. In [13], similar user clustering and beamforming schemes to those in [10]–[12] were applied to multi-cell networks, and the problem of power minimization was tackled jointly with user clustering and beamforming using an iterative distributed methodology. However, the problems of imperfect beam design and the ICI at weak users have not been considered.

In this work, for the sake of eliminating the ICI at weak users in MU-MISO-NOMA system, we propose a new receiver equalizer, namely weak user beam-matching (WBM) equalizer. The proposed equalizer is similar to the transmit beam-matching (TBM) proposed in [14]. The difference between them is, the TBM is used to remove the multiuser interference caused by quantization errors in the case of limited feedback by matching the effective channel with the designed BF vector by per-user unitary rate control (PU^{2}RC) technique, whereas, the WBM equalizer is proposed here to eliminate the ICI at each weak user in MU-MISO-NOMA system by matching its channel with the designed BF vector by ZFBF. The process of beam-matching with WBM equalizer is executed with the aid of D2D CSI sharing, which is a kind of user cooperation that allows nearby users to share their CSI with each other via short-range communications [15]–[17]. In [15], user cooperation is proposed to improve channel feedback accuracy. In [16], the authors proposed a cooperative precoder feedback strategy based on limited CSI sharing via D2D communication, to increase the throughput in a massive MIMO downlink system. In [17], the bi-directional cooperative NOMA was proposed, in which NOMA users cooperate with each other by the channel information exchange to achieve a higher power gain over uni-directional cooperative NOMA and a diversity gain over non-cooperative NOMA and OMA schemes.

To the best of our knowledge, exploiting user cooperation to match the channels of weak users to remove the ICI in a clustered MU-MISO-NOMA system has not been considered in the previous studies. The main contributions of this article can be summarized as follows:

- We propose WBM equalizer at the receivers of weak users to cancel the generated ICI in case of perfect CSIT and limited feedback. The D2D CSI sharing between the nearby strong and weak users of each cluster is exploited to accomplish the process of WBM equalization. With WBM, the ICI can be totally eliminated at weak users.
- The beam-matching concept is also introduced to strong users by proposing strong user beam-matching (SBM) equalizer to eliminate the generated ICI signals at strong users in the case of limited feedback.
- A new dynamic power allocation (PA) strategy is introduced, in which better sum-rate can be achieved for weak users than those in [9]–[12], by considering the gained throughput from removing ICI at weak users.
- In addition to the sum-rate test, the bit error rate (BER) of NOMA users with the proposed equalizers is calculated through designing a full MU-MISO-NOMA system (end-to-end system). The BER[1] performance for these users is also compared with the BER of those in non-cooperative schemes with both perfect and limited feedback scenarios. To the best of our knowledge, a full design for this system is introduced here for the first time in the literature.
- The quantized channels (limited feedback channels) are generated with two models: the conventional random vector quantization (RVQ) and the quantization cell upper bound (QUB) [18], [19]. The quantized channels are used in the sum-rate and BER tests to examine the performance of the network in the case of limited feedback.

The remainder of the paper is organized as follows: Section II describes the system model and the received signals at NOMA users. Section III presents the quantization codebook generation models. Section IV describes the quantized channel decomposition. The proposed beam-matching equalizers are presented in Section V. Section VI presents user clustering. In Section VII, the transmit power allocation is discussed. The End-to-end NOMA system design is introduced in Section VIII. Section IX illustrates the simulation and results. Finally, the paper is concluded in Section X.

A. Notations

Matrices and vectors are denoted by uppercase and lowercase boldface letters, respectively. $\mathbb{E}(\cdot)$ is the expectation of a random variable. $(\cdot)^T$, $(\cdot)^H$, $|\cdot|$ and $\|\cdot\|$ denote the transpose, the Hermitian transpose, the absolute value and the Euclidean norm operators, respectively. Furthermore, important notations are provided in Table I.

TABLE I: SUMMARY OF IMPORTANT NOTATIONS

Notation	Description
$\boldsymbol{h}_k$	Perfect (actual) channel for user k
$\hat{\boldsymbol{h}}_k$	Quantized (limited feedback) channel of $\boldsymbol{h}_k$
$\tilde{\boldsymbol{h}}_k$	Channel direction information $\tilde{\boldsymbol{h}}_k = \boldsymbol{h}_k/\|\boldsymbol{h}_k\|$
$\hat{\boldsymbol{g}}_k$	Quantized channel direction information of $\tilde{\boldsymbol{h}}_k$
γ_k^{we}	CQI of the weak user at the k-th cluster
$\gamma_k^{\overline{we}}$	γ_k^{we} after equalization at weak user side
$\hat{\gamma}^{we}$	Estimation of γ_k^{we} at the BS
$\hat{\gamma}_k^{\overline{we}}$	Estimation of γ_k^{we} after equalization at the BS

II. SYSTEM MODEL

We consider a downlink NOMA with beamforming for multiuser in a single cell network, which consists of a base station (BS) and K randomly distributed single-antenna users, as shown in Fig. 1. The BS is equipped with N_t transmit antennas and can generate N_t BF vectors. Each BF serves two clustered users after superimposing their signals in the power

[1]BER calculation requires a modulation scheme and performing superposition coding for the modulated signal at the transmitter side and implementing SIC at the receiver side to separate the superimposed signals. These processes are not actually required by other tests such as sum-rate and outage probability as they are only based on the SINR expression (SINR is clarified in Section II-C). Thus, BER is considered a more comprehensive test than others.

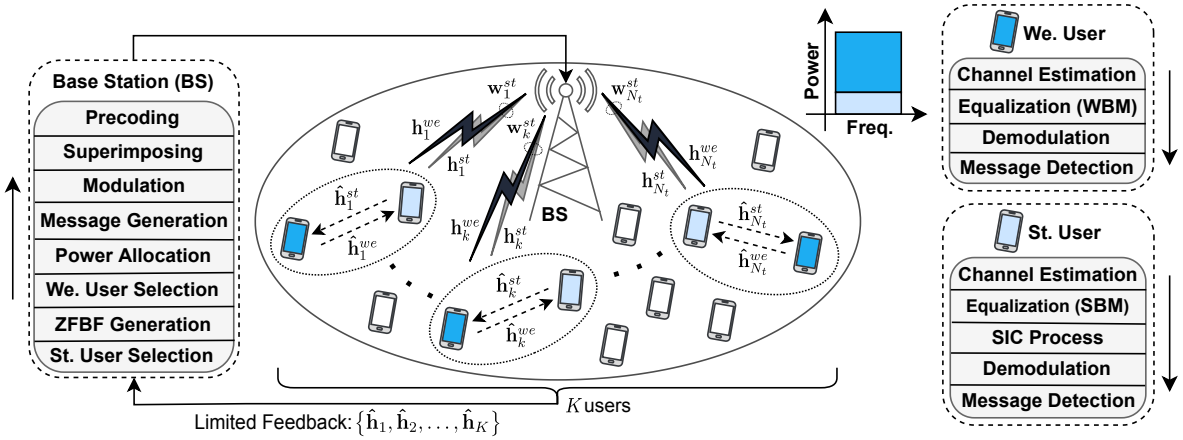


Fig. 1: The proposed end-to-end system model of multiuser cooperative NOMA with D2D-based limited CSI sharing.

domain and precoding the superimposed signal by a single BF vector. The clustered users have different channel qualities, the user with a high channel quality is known as the strong user (denoted by "st"), and the other with a lower channel quality is known as the weak user (denoted by "we"). The strong user removes the signal of the weak user using SIC, while the weak user decodes the received signal directly without SIC.

Throughout this paper, three assumptions are adopted for CSI: *First*, the channel estimation is perfect at each user, and hence perfect CSI is available at receivers (perfect CSIR). *Second*, the CSI feedback to the transmitter has two scenarios: perfect CSIT and limited feedback. *Third*, the CSIs can be shared between strong and weak users within each cluster by exploiting the existing D2D short-range communication technologies that are not overlapping with the cellular spectrum such as Bluetooth, Wi-Fi Direct, or ZigBee [20].

In a basic NOMA without BF system, only superposition coding is applied at the BS. However, in NOMA with ZFBF, the BS firstly superimposes the symbols of each cluster k as

$$x_k = \sqrt{\alpha_k} s_k^{st} + \sqrt{1-\alpha_k} s_k^{we}, \tag{1}$$

where α_k and $(1-\alpha_k)$ are the PA coefficients of the strong and weak users, respectively, such that $(1-\alpha_k) \geq \alpha_k$, s_k^{st} and s_k^{we} are the symbols of the strong and weak users in cluster k, respectively.

After superimposing, the superimposed symbols are saved in $\boldsymbol{x}_s$ vector as,

$$\boldsymbol{x}_s = [x_1, \ldots, x_k, \ldots, x_{N_t}], \quad \boldsymbol{x}_s \in \mathbb{C}^{N_t \times 1} \tag{2}$$

The precoding is then performed on the superimposed symbol vector to generate the superimposed precoded vector $\boldsymbol{x}_s^{Pre}$ as

$$\boldsymbol{x}_s^{Pre} = \boldsymbol{W}^{st} \boldsymbol{x}_s, \quad \boldsymbol{x}_s^{Pre} \in \mathbb{C}^{N_t \times 1} \tag{3}$$

where $\boldsymbol{W}^{st} \in \mathbb{C}^{N_t \times N_t}$ is the ZFBF precoding matrix which is designed based on strong users' channels.

After transmitting the superimposed precoded symbols with total power P for each cluster, the received signals at the strong and weak users of cluster k are respectively given by

$$y_k^{st} = \sqrt{P} \boldsymbol{h}_k^{st} \boldsymbol{w}_k^{st} x_k + \sqrt{P} \sum_{j \neq k} \boldsymbol{h}_k^{st} \boldsymbol{w}_j^{st} x_j + n_k^{st}. \tag{4a}$$

$$y_k^{we} = \sqrt{P} \boldsymbol{h}_k^{we} \boldsymbol{w}_k^{st} x_k + \sqrt{P} \sum_{j \neq k} \boldsymbol{h}_k^{we} \boldsymbol{w}_j^{st} x_j + n_k^{we}. \tag{4b}$$

where $\boldsymbol{w}_k^{st} \in \mathbb{C}^{N_t \times 1}$ is the ZFBF vector of cluster k, $\boldsymbol{h}_k^{st}$, $\boldsymbol{h}_k^{we} \in \mathbb{C}^{1 \times N_t}$ denote the channel vectors of the strong and weak users, respectively. The communication channels from the BS to the mobile users are modeled as uncorrelated Rayleigh flat fading channels with zero mean and unit variance. n_k^{st} and n_k^{we} are the additive white Gaussian noise (AWGN) at these users with distribution $\mathcal{CN}(0, \sigma_n^2)$.

A. Zero-Forcing Beamforming (ZFBF)

To generate ZFBF matrix based on perfect CSIT knowledge of strong users' channels, let

$$\boldsymbol{H}(S_{st}) = \left[(\boldsymbol{h}_1^{st})^T, \cdots, (\boldsymbol{h}_k^{st})^T, \cdots, (\boldsymbol{h}_{N_t}^{st})^T\right]^T, \tag{5}$$

denotes the channel vectors of strong users, where S_{st} denotes the strong users' set. The BS firstly designs the unnormalized precoding matrix $\boldsymbol{W}_0^{st}$ based on $\boldsymbol{H}(S_{st})$ as

$$\begin{aligned} \boldsymbol{W}_0^{st} &= \boldsymbol{H}(S_{st})^H \left(\boldsymbol{H}(S_{st}) \boldsymbol{H}(S_{st})^H\right)^{-1} \\ &= \left[\boldsymbol{w}_{0_1}^{st}, \ldots, \boldsymbol{w}_{0_k}^{st}, \ldots, \boldsymbol{w}_{0_{N_t}}^{st}\right], \end{aligned} \tag{6}$$

Each k-th column of $\boldsymbol{W}_0^{st}$ is then normalized as, $\boldsymbol{w}_k^{st} = \boldsymbol{w}_{0_k} / \|\boldsymbol{w}_{0_k}\|$ to obtain the normalized precoding matrix

$$\boldsymbol{W}^{st} = \left[\boldsymbol{w}_1^{st}, \ldots, \boldsymbol{w}_k^{st}, \ldots, \boldsymbol{w}_{N_t}^{st}\right], \quad \boldsymbol{W}^{st} \in \mathbb{C}^{N_t \times N_t} \tag{7}$$

The precoding matrix $\boldsymbol{W}^{st}$ is then used in the selection process of N_t weak users to make NOMA clusters, as will be described in Section VI.

B. Limited Feedback Channel Model

To cluster users into groups and to perform ZFBF, the BS needs to know the CSI of all users at each time slot. Accordingly, each user needs to estimate its channel based on the received pilots from the BS. After channel estimation, each user feeds back two types of information to the BS, the channel

quality indicator (CQI) and the channel direction information (CDI). The CQI is an estimate of the user's channel quality, and the CDI describes the orientation of the user's channel. For CDI, each user quantizes and feeds back its normalized channel vector denoted by $\tilde{\boldsymbol{h}}_k \triangleq \boldsymbol{h}_k/\|\boldsymbol{h}_k\|$ with B bits. These bits are an index of $N_t \times 1$ BF vector (codeword), which is included into a codebook $\boldsymbol{Q} \triangleq \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_L\}$ that contains $L = 2^B$ independent and isotropically distributed codeword vectors. The index of quantized CDI channel vector $\hat{\boldsymbol{g}}_k$ is chosen from $\boldsymbol{Q}$ according to the following decision rule:

$$\hat{\boldsymbol{g}}_k = \arg\max_{j\in\boldsymbol{Q}} \cos^2 \angle(\tilde{\boldsymbol{h}}_k, \boldsymbol{c}_j) = \arg\min_{j\in\boldsymbol{Q}} \sin^2 \angle(\tilde{\boldsymbol{h}}_k, \boldsymbol{c}_j), \quad (8)$$

Note that the quantization codebook $\boldsymbol{Q}$ is fixed beforehand and known to both, transmitter and mobile users. Therefore, only the index j needs to be fed back to the BS.

For CQI, each mobile user feeds back a single real number represents its channel quality (i.e., channel norm $\|\boldsymbol{h}_k\|$). Prior works have shown that the quantized CQI is virtually the same as the unquantized CQI. Therefore, we assume that the CQI is perfectly available at the BS [16], [19]. Based on the CDI and CQI, the quantized channel for any k-user can be expressed as [16]

$$\hat{\boldsymbol{h}}_k = \|\boldsymbol{h}_k\|\, \hat{\boldsymbol{g}}_k, \quad (9)$$

C. Received Signal Model: Strong User

By substituting (1) in (4a), we can rewrite y_k^{st} before performing the SIC process as

$$y_k^{st} = \underbrace{\sqrt{\alpha_k P}\boldsymbol{h}_k^{st}\boldsymbol{w}_k^{st}s_k^{st}}_{\text{Desired signal}} + \underbrace{\sqrt{(1-\alpha_k)P}\boldsymbol{h}_k^{st}\boldsymbol{w}_k^{st}s_k^{we}}_{\text{Inter-user interference (IUI)}} + \underbrace{\sqrt{P}\sum_{j\neq k}\boldsymbol{h}_k^{st}\boldsymbol{w}_j^{st}x_j}_{\text{Inter-cluster interference (ICI)}} + \underbrace{n_k^{st}}_{\text{Noise}}. \quad (10)$$

After SIC, the IUI caused by weak user can be removed. Also, in the case of perfect CSIT, the ICI is precancelled at the BS. Thus, $\boldsymbol{h}_k^{st}\boldsymbol{w}_j^{st} = 0, \forall j \neq k$. Accordingly, the signal to interference plus noise ratio (SINR) seen by the strong user without these interferences is given by

$$SINR_k^{st} = \frac{\alpha_k P\left|\boldsymbol{h}_k^{st}\boldsymbol{w}_k^{st}\right|^2}{\sigma_n^2}, \quad (11)$$

In the case of limited feedback, the ICI cannot be totally removed, because the precoding matrix is generated based on quantized channels, i.e.,

$$\boldsymbol{W}_0^{st} = \hat{\boldsymbol{H}}\left(S_{st}\right)^H\left(\hat{\boldsymbol{H}}\left(S_{st}\right)\hat{\boldsymbol{H}}\left(S_{st}\right)^H\right)^{-1}, \quad (12)$$

Therefore, $\boldsymbol{h}_k^{st}\boldsymbol{w}_j^{st} \neq 0, \forall j \neq k$ and hence, $SINR_k^{st}$ becomes

$$\begin{aligned} SINR_k^{st} &= \frac{\alpha_k P\left|\boldsymbol{h}_k^{st}\boldsymbol{w}_k^{st}\right|^2}{P\sum_{j\neq k}\left|\boldsymbol{h}_k^{st}\boldsymbol{w}_j^{st}\right|^2 + \sigma_n^2} \\ &= \frac{\alpha_k \rho\left|\boldsymbol{h}_k^{st}\boldsymbol{w}_k^{st}\right|^2}{\rho\sum_{j\neq k}\left|\boldsymbol{h}_k^{st}\boldsymbol{w}_j^{st}\right|^2 + 1} \end{aligned} \quad (13)$$

where $\rho = P/\sigma_n^2$ is the signal-to-noise ratio (SNR) for the k-th cluster. The achievable rate of the strong user is then given by

$$R_k^{st} = \log_2\left(1 + SINR_k^{st}\right). \quad (14)$$

Note that the CQI for the strong user can also be given by

$$\gamma_k^{st} = \frac{\rho\left|\boldsymbol{h}_k^{st}\boldsymbol{w}_k^{st}\right|^2}{\rho\sum_{j\neq k}\left|\boldsymbol{h}_k^{st}\boldsymbol{w}_j^{st}\right|^2 + 1} \quad (15)$$

Based on CQI of the strong user γ_k^{st}, we can rewrite R_k^{st} as

$$R_k^{st} = \log_2\left(1 + \alpha_k\gamma_k^{st}\right). \quad (16)$$

D. Received Signal Model: Weak User

After substituting (1) in (4b), we can rewrite y_k^{we} as

$$y_k^{we} = \sqrt{(1-\alpha_k)P}\boldsymbol{h}_k^{we}\boldsymbol{w}_k^{st}s_k^{we} + \sqrt{\alpha_k P}\boldsymbol{h}_k^{we}\boldsymbol{w}_k^{st}s_k^{st} + \sqrt{P}\sum_{j\neq k}\boldsymbol{h}_k^{we}\boldsymbol{w}_j^{st}x_j + n_k^{we}. \quad (17)$$

Since strong user's BF vector is also applied to weak user, the ICI cannot be totally removed at weak user. Besides, the IUI also appears at weak user, because SIC is not performed by weak user. Therefore, the SINR of weak user contains both IUI and ICI terms and is given by

$$SINR_k^{we} = \frac{(1-\alpha_k)P\left|\boldsymbol{h}_k^{we}\boldsymbol{w}_k^{st}\right|^2}{\alpha_k P\left|\boldsymbol{h}_k^{we}\boldsymbol{w}_k^{st}\right|^2 + P\sum_{j\neq k}\left|\boldsymbol{h}_k^{we}\boldsymbol{w}_j^{st}\right|^2 + \sigma_n^2} \quad (18)$$

Thus, the achievable rate of the weak user is given by

$$\begin{aligned} R_k^{we} &= \log_2\left(1 + SINR_k^{we}\right). \\ &= \log_2\left(1 + \frac{(1-\alpha_k)\gamma_k^{we}}{\alpha_k\gamma_k^{we} + 1}\right). \end{aligned} \quad (19)$$

where

$$\gamma_k^{we} = \frac{\rho\left|\boldsymbol{h}_k^{we}\boldsymbol{w}_k^{st}\right|^2}{\rho\sum_{j\neq k}\left|\boldsymbol{h}_k^{we}\boldsymbol{w}_j^{st}\right|^2 + 1} \quad (20)$$

In the case of limited feedback, the SINRs of weak users becomes even worse, since the ICI is increased due to quantization errors. To overcome this ICI, the BS allocates more power to weak users to increase their SINRs. However, when the number of transmit antennas $N_t > 2$, the SINR of each weak user is highly degraded even if all the transmitted power is allocated to the weak user, because the ICI becomes much higher when N_t increases.

Therefore, in this work, WBM equalizer is proposed at the receivers of weak users to cancel out this interference. Please refer to Section V-A for more details.

III. Quantization Codebook Generation Models

The optimal codebook design is not generally known and it depends on various specifics of the system [21]. Besides, the constructing complexity of a codebook $\boldsymbol{Q}$ is highly increased with the number of feedback bits B. Therefore, we resort to quantization cell approximation models to estimate the effect of limited feedback channels. These models are: The conventional random vector quantization (RVQ) [22] and the

quantization cell upper bound (QUB) [18], [19]. With these models, the quantized channel $\hat{\boldsymbol{g}}_k$ can be effectively generated with less complexity than that generated using (8) [22].

These models are based on the ideal assumption that each quantization cell is a Voronoi region around a quantization vector (two-sided) of a spherical cap with the surface area 2^{-B} of the total surface area of the unit sphere [19] as illustrated in Fig. 2. Where θ_k in Fig. 2 represents the error angle between the actual channel direction $\tilde{\boldsymbol{h}}_k$ and its quantization $\hat{\boldsymbol{g}}_k$.

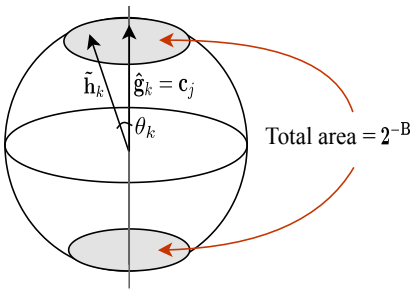


Fig. 2: Quantization cell approximation denoted by the shaded regions.

A. Random Vector Quantization (RVQ) Model

Let us firstly review some basics of RVQ that will be useful in the later derivations. As stated earlier, the 2^B codeword vectors are independently distributed into N_t-dimensional unit sphere codebooks $\boldsymbol{Q}$. In order to determine the quantization error, which is the most important quantity in limited feedback systems, let we first consider the inner product between a channel direction and an arbitrary codeword vector

$$\cos^2 \angle(\tilde{\boldsymbol{h}}_k, \boldsymbol{c}_j) = |\tilde{\boldsymbol{h}}_k \boldsymbol{c}_j|^2, \tag{21}$$

Because $\tilde{\boldsymbol{h}}_k$ and $\boldsymbol{c}_j$ are independent isotropic vectors, the quantity in (21) is beta distributed with parameters $(1,\ N_t-1)$, and

$$\sin^2 \theta_k = \sin^2 \angle(\tilde{\boldsymbol{h}}_k, \boldsymbol{c}_j) = 1 - \cos^2 \angle(\tilde{\boldsymbol{h}}_k, \boldsymbol{c}_j), \tag{22}$$

is the quantization error between real and quantized channel which also has a beta distribution with parameters $(N_t-1,\ 1)$.

Since $\hat{\boldsymbol{g}}_k$ is the quantization of the vector $\tilde{\boldsymbol{h}}_k$, i.e., the solution of (8) and these vectors are independent, the quantization error $\sin^2 \theta_k$ is the minimum of 2^B, and the complementary cumulative distribution function (CCDF) distribution function of $\sin^2 \theta_k$ is given by [23, Lemma 1]

$$F_{\sin^2 \theta_k}(z) = \Pr\left(\sin^2 \theta_k \le z\right) = (1-z^{N_t-1})^{2^B}, \quad 0 \le z \le 1 \tag{23}$$

The expectation of $\sin^2 \theta_k$ is given by

$$\mathbb{E}[\sin^2 \theta_k] = \int_0^1 \left(1 - z^{N_t-1}\right)^{2^B} dz, \tag{24}$$

The result can alternatively be derived using an integral representation of the beta function as [22], [24]

$$\beta\big(c, \frac{a}{b}\big) = b \int_0^1 z^{a-1}\big(1 - z^b\big)^{c-1}\, dz,\ a > 0,\ b > 0,\ c > 0$$

Substituting $a = 1,\ b = N_t - 1$ and $c = \left(2^B + 1\right)$ we get

$$\beta\left(2^B + 1, \frac{1}{N_t - 1}\right) = (N_t - 1) \int_0^1 (1 - z^{N_t-1})^{2^B}\, dz, \tag{25}$$

From (24) and (25), we get

$$\begin{aligned}
\mathbb{E}[\sin^2 \theta_k] &= 2^B \beta\left(2^B, \frac{N_t}{N_t - 1}\right) \\
&= \frac{1}{N_t - 1} \beta\left(2^B + 1, \frac{1}{N_t - 1}\right), \\
&= \frac{\frac{1}{N_t-1}\Gamma\left(2^B + 1\right) + \Gamma\left(\frac{1}{N_t-1}\right)}{\Gamma\left(2^B + 1 + \frac{1}{N_t-1}\right)} \\
&= \frac{2^B\Gamma\left(2^B\right) + \Gamma\left(1 + \frac{1}{N_t-1}\right)}{\Gamma\left(2^B + 1 + \frac{1}{N_t-1}\right)} \\
&= 2^B \cdot \beta\left(2^B, \frac{N_t}{N_t - 1}\right),
\end{aligned} \tag{26}$$

where $\Gamma(\cdot)$ is the Gamma function. On the other hand, we have used the fundamental equality $\Gamma(\tau + 1) = \tau\Gamma(\tau)$. Here, $\beta(\cdot)$ is used to denote the beta function, which is defined in terms of the gamma function as $\beta(\tau, u) = \frac{\Gamma(\tau)\Gamma(u)}{\Gamma(\tau+u)}$ [22]. Hence, for $N_t > 2$, the result in (26) will be

$$2^B \cdot \beta\left(2^B, \frac{N_t}{N_t - 1}\right) \le \frac{2^B\Gamma\left(2^B\right)}{\Gamma\left(2^B + 1 + \frac{1}{N_t-1}\right)}$$

The preceding inequality is reached because $\Gamma(\tau) \le 1$ for $1 \le \tau \le 2$, due to the convexity of the gamma function and the fact that $\Gamma(1) = \Gamma(2) = 1$. By applying Kershaw's inequality for the gamma function [22]

$$\frac{\Gamma(\tau + m)}{\Gamma(\tau + 1)} < \left(\tau + \frac{m}{2}\right)^{m-1}, \forall \tau > 0,\ 0 < m < 1$$

with $\tau = 2^B + \frac{1}{N_t-1}$ and $m = 1 - \frac{1}{N_t-1}$, we get

$$\frac{\Gamma\left(2^B + 1\right)}{\Gamma\left(2^B + 1 + \frac{1}{N_t-1}\right)} < \left(2^B + \frac{N_t}{2(N_t - 1)}\right)^{-\frac{1}{N_t-1}}$$

With the decreasing nature of $(\cdot)^{-\frac{1}{N_t-1}}$ function, the expected quantization error can be upper-bounded as

$$\mathbb{E}\left[\sin^2 \theta_k\right] < 2^{-\frac{B}{N_t-1}} \tag{27}$$

Note that this upper bound can be represented in Fig. 2, when θ_k lies between the center and a vector on the boundary of the spherical cap, i.e., $\sin^2 \theta_k = \sin^2 \angle(\tilde{\boldsymbol{h}}_k, \hat{\boldsymbol{g}}_k) = 2^{-\frac{B}{N_t-1}}$.

Without loss of generality, to quantify the quantization error within $2^{-\frac{B}{N_t-1}}$ limit, the cumulative distribution function (CDF) in (23) can be rewritten with a new upper bound for the quantization error as [22]

$$F_{\sin^2 \theta_k}(z) = 1 - z^{N_t-1}, \qquad 0 \le z < \left(1 - v^{\frac{1}{L}}\right)^{\frac{1}{N_t-1}} \tag{28}$$

where the $(1 - v^{\frac{1}{L}})^{1/(N_t-1)}$ term, is the upper bound limit of a quantization error and v is a random variable within the

interval $[0,1]$. Therefore, the final closed-form of quantization error based on RVQ is given by

$$\sin^2\theta_k = \left(1 - v^{\frac{1}{L}}\right)^{\frac{1}{N_t-1}}, \text{ using } \quad v \in [0,1] \tag{29}$$

It is important to mention that RVQ is not an optimal quantization approach, because the quantization error generated by (29) is not always $< 2^{-\frac{B}{N_t-1}}$. However, the results in Section IX indicate that the RVQ model is still accurate to characterize the degradation in the performance of throughput and BER in the case of limited feedback.

B. Quantization Cell Upper Bound (QUB) Model

To derive the CDF of QUB, we start from the geometrical framework introduced in [25]. Around each codebook vector $\boldsymbol{c}_j$, the surface area of the spherical cap is defined as

$$\Psi_j(z) = \left\{\tilde{\boldsymbol{h}}_k : \sin^2\angle(\tilde{\boldsymbol{h}}_k, \boldsymbol{c}_j) \le z\right\}, \ 0 \le z \le 1 \tag{30}$$

Let $A\{\Psi_j(z)\}$ denotes the area of the spherical cap $\Psi_j(z)$, and $A\{\Psi_j(1)\}$ is the whole surface area of the unit hypersphere. According to [25, Lemma 2], the surface area of the spherical cap $\Psi_j(z)$ is

$$A\big\{\Psi_j(z)\big\} = \frac{2\pi^{N_t} z^{N_t-1}}{(N_t-1)!} \tag{31}$$

Note that

$$F_{\sin^2\theta_k}(z) = \Pr\left\{\arg\min_{j\in\boldsymbol{Q}} \sin^2\angle(\tilde{\boldsymbol{h}}_k, \boldsymbol{c}_j) \le z\right\} \tag{32}$$

can be interpreted as the probability that $\tilde{\boldsymbol{h}}_k$ falls in the union of the regions $\big\{\Psi_j(z)\big\}_{j=1}^{L}$ (denoted by $\cup_{j=1}^{L}\Psi_j(z)$). Since $\boldsymbol{h}_k$ is i.i.d [25], and $\tilde{\boldsymbol{h}}_k$ is uniformly distributed on the surface of the unit sphere, $F_{\sin^2\theta_k}(z)$ can be represented as [18], [25]

$$F_{\sin^2\theta_k}(z) = \frac{A\big\{\cup_{j=1}^{L}\Psi_j(z)\big\}}{A\big\{\Psi_j(1)\big\}} \tag{33}$$

Due to the union operation, it is true that

$$A\big\{\cup_{j=1}^{L}\Psi_j(z)\big\} \le \sum_{j=1}^{L} A\big\{\Psi_j(1)\big\} \tag{34}$$

From area computation given in (31), we obtain

$$F_{\sin^2\theta_k}(z) = \frac{A\big\{\cup_{j=1}^{L}\Psi_j(z)\big\}}{A\big\{\Psi_j(1)\big\}} = Lz^{N_t-1} \tag{35}$$

Taking into account, that $F_{\sin^2\theta_k}(z) \le 1$, the CDF can be an upper bounded as [18], [19]

$$F_{\sin^2\theta_k}(z) = \begin{cases} Lz^{N_t-1}, & 0 \le z < \left(\frac{1}{L}\right)^{\frac{1}{N_t-1}}, \\ 1, & z \ge \left(\frac{1}{L}\right)^{\frac{1}{N_t-1}}, \end{cases} \tag{36}$$

Therefore, to generate a quantization error $\sin^2\theta_k < 2^{-\frac{B}{N_t-1}}$ based on QUB, we should have

$$\sin^2\theta_k = \left(\frac{v}{L}\right)^{\frac{1}{N_t-1}}, \text{ using } \quad v \in [0,1] \tag{37}$$

In this way, $\sin^2\theta_k$ will always give values lower than $2^{-\frac{B}{N_t-1}}$.

IV. Quantized Channel Decomposition

Based on the magnitude of the quantization error $\sin^2\theta_k$, the quantized channel direction $\hat{\boldsymbol{g}}_k$ can be decomposed as a summation of two components, one in the direction of the quantization, and the other is isotropically distributed in the nullspace (orthogonal) of the quantization as follows:

$$\hat{\boldsymbol{g}}_k = \cos\theta_k\tilde{\boldsymbol{h}}_k + \sin\theta_k\tilde{\boldsymbol{e}}_k, \tag{38}$$

where the error angle $\theta_k \in [0, \pi/2]$ and $\tilde{\boldsymbol{e}}_k = \boldsymbol{e}_k/\|\boldsymbol{e}_k\|$ is a unit norm vector isotropically distributed in the nullspace of $\tilde{\boldsymbol{h}}_k$ and is independent of $\sin^2\theta_k$.

Note that the quantized channel given in (38), gives a normalized quantized channel. To remove the effect of normalization, $\hat{\boldsymbol{g}}_k$ is multiplied by the CQI-based channel norm $\|\boldsymbol{h}_k\|$ to generate $\hat{\boldsymbol{h}}_k$ as [16], [19]

$$\hat{\boldsymbol{h}}_k = \|\boldsymbol{h}_k\|\left(\cos\theta_k\tilde{\boldsymbol{h}}_k + \sin\theta_k\tilde{\boldsymbol{e}}_k\right), \tag{39}$$

where $\hat{\boldsymbol{h}}_k$ represents an alternative quantized channel to that given in (9), and it is generated based on QUB model.

In Algorithm 1, we summarize the required steps to generate the quantized channels for all users K based on QUB model. It is worth to mention here that Algorithm 1 can also be used to generate quantized channels based on RVQ model using (29) instead of (37).

Algorithm 1 Quantized Channel Generation based on QUB and Channel Norm

for all $k = 1$ to K **do**

1) Generate a quantization error for user k using (37).
2) Determine the orthogonal error vector $\boldsymbol{e}_k$ as,
$$\boldsymbol{e}_k = \left(\boldsymbol{r} - \tilde{\boldsymbol{h}}_k\boldsymbol{r}\tilde{\boldsymbol{h}}_k^H\right)^H, \tag{40}$$
where $\boldsymbol{r}$ is a random variable $\boldsymbol{r} \in \mathbb{C}^{N_t\times 1}$.
3) Normalize $\tilde{\boldsymbol{e}}_k = \frac{\boldsymbol{e}_k}{\|\boldsymbol{e}_k\|}$,
4) Calculate $\hat{\boldsymbol{h}}_k$ as,
$$\hat{\boldsymbol{h}}_k = \|\boldsymbol{h}_k\|\left(\cos\theta_k\tilde{\boldsymbol{h}}_k + \sin\theta_k\tilde{\boldsymbol{e}}_k\right),$$

end for

V. The Proposed beam-matching Equalizers

In downlink OMA systems such as orthogonal frequency division multiple access (OFDMA), the interference is avoided by assigning orthogonal subcarriers to different users at the expense of dividing the whole bandwidth between subcarriers. In time division multiple access (TDMA) system, a nonoverlapping time slot is assigned for each active user to avoid the interference. In MU-MISO-NOMA system, the bandwidth and time resources are not divided and are fully available to each of the clustered users. The ICI is precancelled through ZFBF precoding, which is also known as space division multiple access (SDMA). However, the ICI can only be removed at strong users, while weak users suffer from a residual ICI due to BF mismatch effect. Therefore, we propose weak user beam-matching (WBM) equalizer at weak users' sides to remove this ICI. In addition, strong user beam-matching (SBM) equalizer

is proposed for strong users to eliminate the generated ICI in the case of limited feedback. In the following subsections, more details are provided about these equalizers.

A. Weak user Beam-Matching (WBM) Equalizer

WBM equalizer is proposed here as a post-processing step at weak user to match its channel with the designed BF vector. The equalization of WBM relies on CSI sharing between the strong and weak user of each cluster to obtain the CSI of the strong user and use it in the process of beam-matching.

Without loss of generality, it is assumed that the nearby strong and weak users within each cluster k can cooperatively share their CSIs with each other, as shown in Fig. 1. Therefore, after CSI sharing, each weak user can equalize its received signal as follows

$$\begin{aligned} y_k^{\overline{we}} &= \frac{\boldsymbol{h}_k^{st}}{\boldsymbol{h}_k^{we}} y_k^{we} \\ y_k^{\overline{we}} &= \frac{\boldsymbol{h}_k^{st}}{\boldsymbol{h}_k^{we}} \Big[\sqrt{(1-\alpha_k)\,P}\boldsymbol{h}_k^{we}\boldsymbol{w}_k^{st} s_k^{we} + \sqrt{\alpha_k P}\boldsymbol{h}_k^{we}\boldsymbol{w}_k^{st} s_k^{st} \\ &\quad + \sqrt{P}\sum_{j\neq k} \boldsymbol{h}_k^{we}\boldsymbol{w}_j^{st} x_j + n_k^{we} \Big]. \end{aligned} \tag{41}$$

where the ratio $\boldsymbol{h}_k^{st}/\boldsymbol{h}_k^{we} \in \mathbb{C}^{1\times 1}$ is the WBM equalization factor for the case of perfect CSI sharing. This factor is multiplied by the received signal y_k^{we} to remove the effect of weak user's channel by replacing it with the strong user's channel. After WBM equalization, we can rewrite (41) as

$$\begin{aligned} y_k^{\overline{we}} &= \sqrt{(1-\alpha_k)\,P}\boldsymbol{h}_k^{st}\boldsymbol{w}_k^{st} s_k^{we} + \sqrt{\alpha_k P}\boldsymbol{h}_k^{st}\boldsymbol{w}_k^{st} s_k^{st} \\ &\quad + \sqrt{P}\sum_{j\neq k} \boldsymbol{h}_k^{st}\boldsymbol{w}_j^{st} x_j + \frac{\boldsymbol{h}_k^{st}}{\boldsymbol{h}_k^{we}} n_k^{we}. \end{aligned} \tag{42}$$

Since $\boldsymbol{h}_k^{st}\boldsymbol{w}_j^{st} = 0, \forall j \neq k$, can be achieved after WBM equalization, the ICI can be totally removed at weak users and therefore, $y_k^{\overline{we}}$ in (42) reduces to

$$y_k^{\overline{we}} = \sqrt{(1-\alpha_k)\,P}\boldsymbol{h}_k^{st}\boldsymbol{w}_k^{st} s_k^{we} + \sqrt{\alpha_k P}\boldsymbol{h}_k^{st}\boldsymbol{w}_k^{st} s_k^{st} + \frac{\boldsymbol{h}_k^{st}}{\boldsymbol{h}_k^{we}} n_k^{we}. \tag{43}$$

Accordingly, the $SINR_k^{\overline{we}}$ after WBM equalization becomes

$$SINR_k^{\overline{we}} = \frac{(1-\alpha_k)\,P\,|\boldsymbol{h}_k^{st}\boldsymbol{w}_k^{st}|^2}{\alpha_k P\,|\boldsymbol{h}_k^{st}\boldsymbol{w}_k^{st}|^2 + \left|\frac{\boldsymbol{h}_k^{st}}{\boldsymbol{h}_k^{we}}\right|^2 \sigma_n^2} \tag{44}$$

In the case of limited feedback, the users share their quantized channels. Thus, the received signal after equalization becomes

$$y_k^{\overline{we}} = \frac{\hat{\boldsymbol{h}}_k^{st}}{\boldsymbol{h}_k^{we}} y_k^{we}, \tag{45}$$

and accordingly, the $SINR_k^{\overline{we}}$ turns to

$$SINR_k^{\overline{we}} = \frac{(1-\alpha_k)\,P|\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_k^{st}|^2}{\alpha_k P|\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_k^{st}|^2 + \left|\frac{\hat{\boldsymbol{h}}_k^{st}}{\boldsymbol{h}_k^{we}}\right|^2 \sigma_n^2} \tag{46}$$

Since $\boldsymbol{W}_0^{st}$ is computed based on the quantized channels, i.e., $\hat{\boldsymbol{H}}\,(S_{st}) = \left[(\hat{\boldsymbol{h}}_1^{st})^T, \cdots, (\hat{\boldsymbol{h}}_k^{st})^T, \cdots, (\hat{\boldsymbol{h}}_{N_t}^{st})^T\right]^T$, we can also have $\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_j^{st} = 0, \forall j \neq k$. Hence, a similar performance can be expected with limited feedback to that with perfect CSIT.

B. Strong user Beam-Matching (SBM) Equalizer

In the case of limited feedback, $\boldsymbol{h}_k^{st}\boldsymbol{w}_j^{st} \neq 0, \forall j \neq k$, since $\boldsymbol{w}_j^{st}$ is fully orthogonal to the quantized channel $\hat{\boldsymbol{h}}_k^{st}$ but not the actual channel $\boldsymbol{h}_k^{st}$. However, this leads to a decrease in the SINR value at strong user due to occurring ICI. Therefore, strong user beam-matching (SBM) equalizer is proposed here to cancel this ICI based on the same principle of WBM equalizer. Accordingly, the received signal y_k^{st} is equalized after multiplying (10) by the SBM equalization factor as follows

$$\begin{aligned} y_k^{\overline{st}} &= \frac{\hat{\boldsymbol{h}}_k^{st}}{\boldsymbol{h}_k^{st}} y_k^{st} \\ y_k^{\overline{st}} &= \frac{\hat{\boldsymbol{h}}_k^{st}}{\boldsymbol{h}_k^{st}} \Big[\sqrt{\alpha_k P}\boldsymbol{h}_k^{st}\boldsymbol{w}_k^{st} s_k^{st} + \sqrt{(1-\alpha_k)\,P}\boldsymbol{h}_k^{st}\boldsymbol{w}_k^{st} s_k^{we} \\ &\quad + \sqrt{P}\sum_{j\neq k} \boldsymbol{h}_k^{st}\boldsymbol{w}_j^{st} x_j + n_k^{st} \Big] \\ &= \sqrt{\alpha_k P}\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_k^{st} s_k^{st} + \sqrt{(1-\alpha_k)\,P}\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_k^{st} s_k^{we} \\ &\quad + \sqrt{P}\sum_{j\neq k} \hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_j^{st} x_j + \frac{\hat{\boldsymbol{h}}_k^{st}}{\boldsymbol{h}_k^{st}} n_k^{st}. \end{aligned} \tag{47}$$

After SBM equalization, we can have $\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_j^{st} = 0, \forall j \neq k$, as $\hat{\boldsymbol{h}}_k^{st}$ is completely orthogonal to $\boldsymbol{w}_j^{st}$. Thus, the ICI can be nulled[2] at strong user.

After performing SIC process, the $\sqrt{(1-\alpha_k)\,P}\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_k^{st} s_k^{we}$ term is also removed. Hence, we can rewrite (47) as

$$y_k^{\overline{st}} = \sqrt{\alpha_k P}\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_k^{st} s_k^{st} + \frac{\hat{\boldsymbol{h}}_k^{st}}{\boldsymbol{h}_k^{st}} n_k^{st}. \tag{48}$$

Correspondingly, the $SINR_k^{st}$ after SBM and SIC processes can be rewritten as

$$SINR_k^{\overline{st}} = \frac{\alpha_k P|\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_k^{st}|^2}{\left|\frac{\hat{\boldsymbol{h}}_k^{st}}{\boldsymbol{h}_k^{st}}\right|^2 \sigma_n^2} \tag{49}$$

C. Complexity Analysis of WBM and SBM Equalizers

The use of WBM or SBM equalizer adds a small complexity to the process of signal detection at each user. For example, the equalization factor of SBM equalizer $\hat{\boldsymbol{h}}_k^{st}/\boldsymbol{h}_k^{st}$, needs: $(1\times N_t)/(1\times N_t) = (1\times N_t)\times(N_t\times 1)$ vector-matrix multiplication[3], which takes $2N_t$ real multiplications and $2N_t - 1$ real additions [12]. Therefore, the total flop count is $4N_t - 1$. Hence, the added complexity from SBM equalizer is only $\mathcal{O}(N_t)$, which is the same as that added from WBM equalizer.

[2]Even with the general assumption that the perfect CSI is originally available at the receiver side (perfect CSIR). In practical, the channel estimation of any receiver is not perfect. For example, with SBM equalizer, $(\boldsymbol{h}_k^{st})_{est}/\boldsymbol{h}_k^{st} \neq 1$ in (47), since the estimated $(\boldsymbol{h}_k^{st})_{est} \approx \boldsymbol{h}_k^{st}$ and this may result in a small interference that can be ignored without affect the main insights of the results.

[3]The division of $\hat{\boldsymbol{h}}_k^{st}/\boldsymbol{h}_k^{st}$ results in a $(1\times N_t)\times(N_t\times 1) = 1\times 1$ scalar quantity, i.e., $\hat{\boldsymbol{h}}_k^{st}/\boldsymbol{h}_k^{st} \in \mathbb{C}^{1\times 1}$, and $y_k^{st} \in \mathbb{C}^{1\times 1}$ is also a scalar quantity. This verifies the multiplication operations between the equalization factors and their corresponding received signals in (41) and (47), respectively, because they have the same dimensions.

VI. User Clustering

In order to enhance the rate performance of MU-MISO-NOMA system, the users should be grouped into clusters. The first clustering algorithm presented in [9] for MU-MISO-NOMA system was based on two factors: a high channel correlation and a large channel gain-difference between the users in a cluster. Despite the superior performance of NOMA system with this algorithm over TDMA system, the weak users suffer from decreasing their sum-rate when the number of users K increases. This is because the probability of choosing the weak user with a smaller channel gain becomes larger when K increases. In [10], this problem was solved by adopting semiorthogonal user selection (SUS) algorithm [26] to select the strong users and maximum signal to interference ratio (SIR) to select weak users. Besides, it has been shown in [12], [27] that using the maximum product of effective channel gains (MPECG) instead of SUS provides better sum-rate especially in the case of limited feedback, Therefore, in this work, we resort to the MPECG-SIR algorithm to cluster NOMA users. With MPECG, the first strong user is selected based on maximum channel gain as

$$S_{st}\left(1\right)=\arg\max_{i\in U}\|\hat{\boldsymbol{h}}_i\|^2,\ U=\left\{1,\ldots,K\right\} \tag{50}$$

The remaining strong users at any cluster k are selected from the remaining user set $C=U-S_{st}\left(1\right)$ as

$$S_{st}\left(k\right)=\arg\max_{i\in C}\prod_{j=1}^{k}\lambda_j, \tag{51}$$

where λ_j is the effective channel gain and is given by

$$\lambda_j=\frac{1}{\left[\hat{\boldsymbol{H}}\left(S_{st}\right)\hat{\boldsymbol{H}}\left(S_{st}\right)^H\right]^{-1}_{j,j}}, \tag{52}$$

and the matrix $\hat{\boldsymbol{H}}\left(S_{st}\right)$ is changed for each user i as

$$\hat{\boldsymbol{H}}\left(S_{st}\right)=\left[\hat{\boldsymbol{H}}\left(S_{st}\left(k-1\right)\right)^T,\hat{\boldsymbol{h}}_i^T\right]^T,i\in C \tag{53}$$

After selecting N_t strong users, ZFBF vectors are generated and normalized as in (7), to be used in the selection process of weak users through SIR scheme [10]. According to the SIR, the selection of a weak user for any cluster k is given by

$$S_{we}\left(k\right)=\arg\max_{i\in C}\left(\frac{|\hat{\boldsymbol{h}}_i\boldsymbol{w}_k^{st}|^2}{\sum_{j\neq k}|\hat{\boldsymbol{h}}_i\boldsymbol{w}_j^{st}|^2}\right),C=U-S_{st} \tag{54}$$

where S_{we} is the weak users' set. The SIR selection is completed after selecting N_t weak users. Therefore, the total number of clustered users with MPECG-SIR algorithm is $2N_t$ users. In (54), it can be noticed that weak user's selection is achieved by the user i that has the highest channel correlation with respect to the strong user, since $\boldsymbol{w}_k^{st}$ is designed based on the strong user's channel. The high channel correlation usually appears between the nearby users, and this allows exploiting the D2D CSI sharing between the clustered users.

It is worth noting here that the number of supportable users with OMA systems (e.g., TDMA system) is equal to N_t, whereas this NOMA scheme can double the number of supported users by OMA systems, and hence it could meet the expected massive connectivity of B5G networks.

VII. The Proposed Power Allocation

After user clustering, the BS should allocate suitable power portions for the users of each cluster to guarantee high data rates [28]. In [10], [12], the PA strategies were adaptive with TDMA system "Adap. NOMA", since NOMA is switched to the TDMA transmission if the PA fails to support the target sum-rate and causes $\alpha_k<0$. However, this failure occurs most likely when N_t increases and causes high ICI $\left(\text{i.e., } P\sum_{j\neq k}|\boldsymbol{h}_k^{we}\boldsymbol{w}_j^{st}|^2\right)$ at weak users.

By relying on D2D CSI sharing and WBM equalizer, the ICI at weak users can be cancelled out, and therefore, a redundant power can be gained and allocated for both strong and weak users to achieve higher sum-rate than without exploiting WBM equalizer. Therefore, we extend the conventional PA strategy [10], [12] by considering the gained throughput from interference cancellation at weak users in our proposed PA strategy. To do so, the sum-rate optimization problem in [10], [12] is reformulated as follows:

$$\max_{\alpha_k}\quad\left(R_k^{we}+R_k^{st}\right) \tag{55a}$$

$$\text{s.t.}\quad R_k^{we}>R_{k-TDMA}^{we}+R_{IC-Gain}, \tag{55b}$$

$$0\leq\alpha_k\leq 1, \tag{55c}$$

$$\left(1-\alpha_k\right)\geq\alpha_k \tag{55d}$$

where

$$R_{k-TDMA}^{we}=\frac{1}{2}\log_2\left(1+\frac{\rho|\hat{\boldsymbol{h}}_k^{we}\boldsymbol{w}_k^{we}|^2}{\rho\sum_{j\neq k}|\hat{\boldsymbol{h}}_k^{we}\boldsymbol{w}_j^{we}|^2+1}\right) \tag{56}$$

is the corresponding sum-rate of the weak user if it would be supported by TDMA, and $R_{IC-Gain}$ denotes the interference cancellation throughput gain and it is given by

$$R_{IC-Gain}=\frac{1}{2}\log_2\left(1+\hat{\gamma}_k^{\overline{we}}-\hat{\gamma}_k^{we}\right), \tag{57}$$

where

$$\hat{\gamma}_k^{\overline{we}}=\frac{\rho|\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_k^{st}|^2}{\rho\sum_{j\neq k}|\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_j^{st}|^2+\left|\frac{\hat{\boldsymbol{h}}_k^{st}}{\hat{\boldsymbol{h}}_k^{we}}\right|^2} \tag{58}$$

denotes the estimated CQI after equalization at the BS for the weak user in the k-th cluster. Note that the $1/2$ in (56) is needed to match TDMA with NOMA system, since TDMA system requires two time slots to support $2N_t$ users. Whereas the $1/2$ in (57) is required to divide $R_{IC-Gain}$ between strong and weak users and for consistency with R_{k-TDMA}^{we}.

It should be clear that, R_k^{we}, R_k^{st}, R_{k-TDMA}^{we}, and $\hat{\gamma}_k^{\overline{we}}$ are computed based on the quantized channels, since limited feedback is assumed to be available at the BS.

The problem defined in (55) is convex with respect to α_k and its Karush-Kuhn-Tucker (KKT) conditions are given as follows:

$$\frac{\partial\left(R_k^{we}+R_k^{st}\right)}{\partial\alpha_k}=\Lambda\frac{\partial\left(R_k^{we}-R_{k-TDMA}^{we}-R_{IC-Gain}\right)}{\partial\alpha_k}, \tag{59a}$$

$$\Lambda\geq 0, \tag{59b}$$

$$R_k^{we}-\left(R_{k-TDMA}^{we}+R_{IC-Gain}\right)\leq 0, \tag{59c}$$

$$\Lambda\left(R_k^{we}-\left(R_{k-TDMA}^{we}+R_{IC-Gain}\right)\right)=0, \tag{59d}$$

where Λ is the Lagrange multipliers for the constraints. Clearly, $\Lambda \neq 0$. Otherwise, $\alpha_k < 0$ does not satisfy (55c). Therefore, we can solve condition (59c) in (59) for the optimal solution as:

$$R_k^{we} = R_{k-TDMA}^{we} + R_{IC-Gain},$$
$$\log_2\left(1+\frac{(1-\alpha_k)\,\hat{\gamma}_k^{\overline{we}}}{\alpha_k\hat{\gamma}_k^{\overline{we}}+1}\right) = \frac{1}{2}\Bigg[\log_2\left(1+SINR_{k-TDMA}^{we}\right) + \log_2\left(1+\hat{\gamma}_k^{\overline{we}}-\hat{\gamma}_k^{we}\right)\Bigg]. \tag{60}$$

Hence, the optimal power fraction of strong user α_k can be obtained from (60), and it is given by:

$$\alpha_k = \frac{1+\frac{1}{\hat{\gamma}_k^{\overline{we}}}}{\sqrt{\left(1+\hat{\gamma}_k^{\overline{we}}-\hat{\gamma}_k^{we}\right)\left(1+SINR_{k-TDMA}^{we}\right)}} - \frac{1}{\hat{\gamma}_k^{\overline{we}}} \tag{61}$$

It is important to mention here, that the proposed NOMA system with this PA strategy is not adaptive with TDMA, because α_k is always positive ($\alpha_k > 0$), as the condition (55b) can always be satisfied.

VIII. The proposed End-to-end System Design

The proposed end-to-end MU-MISO-NOMA system which is illustrated in Fig. 1 is discussed here, starting from signal generation to signal detection at NOMA users. The details of the system design are explained in the following subsections.

A. Superimposed Signals Generation and Modulation

First, the symbols of NOMA users are randomly generated with quadrature phase-shift keying (QPSK) modulation. These symbols are then combined in the power domain using (1) to yield a superimposed NOMA signal. After that, ZFBF vectors are generated based on the channels of strong users and used to precode the superimposed signals of the clustered users using (3). Finally, superimposed signals are transmitted by the BS to the clustered users over Rayleigh channel.

B. Signal Detection

At strong user, the SBM equalization is performed to get the equalized signal $y_k^{\overline{st}}$. Then, the symbol of the weak user is firstly estimated at the strong user before the SIC process as

$$\hat{s}_k^{we_{st}} = \left\langle \frac{y_k^{\overline{st}}}{\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_k^{st}\sqrt{(1-\alpha_k)P}} \right\rangle, \tag{62}$$

where $\langle\cdot\rangle$ denotes signal detection and demodulation.

After that, SIC is performed, by subtracting the estimated weak user interference signal to detect strong user's symbol as

$$\hat{s}_k^{st} = \left\langle \frac{\frac{y_k^{\overline{st}}}{\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_k^{st}} - \sqrt{(1-\alpha_k)P}\,\hat{s}_k^{we_{st}}}{\sqrt{\alpha_k P}} \right\rangle, \tag{63}$$

On the other hand, the estimation of weak user's symbol at the weak user side is performed after WBM equalization as

$$\hat{s}_k^{we_{we}} = \left\langle \frac{y_k^{\overline{we}}}{\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_k^{st}\sqrt{(1-\alpha_k)P}} \right\rangle, \tag{64}$$

Note that in real transmission, the term $\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_k^{st}$ in (62)–(64) should be replaced by $\|\hat{\boldsymbol{h}}_k^{st}\|$, since $\boldsymbol{w}_k^{st}$ is only known to the transmitter, and $\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_k^{st} \approx \|\hat{\boldsymbol{h}}_k^{st}\|$. Also, in (63), the division $y_k^{\overline{st}}/\hat{\boldsymbol{h}}_k^{st}\boldsymbol{w}_k^{st}$ is performed before SIC, because in (1), NOMA PA is implemented at the symbol level and the division on the channel gain returns the received signal to the symbol level, to guarantee a correct implementation of SIC process. Besides, in adaptive NOMA system, to estimate the transmitted symbols properly, the received signals y_k^{st} and y_k^{we} in (62)–(64) should be divided by $\boldsymbol{h}_k^{st}\boldsymbol{w}_k^{st}$ and $\boldsymbol{h}_k^{we}\boldsymbol{w}_k^{st}$, respectively.

In Algorithm 2, the required steps to design an end-to-end MU-MISO-NOMA system are summarized.

Algorithm 2 End-to-end MU-MISO-NOMA System Design Using WBM and SBM Equalizers with Limited Feedback

Step 1) User Clustering & ZFBF Generation

1: Generate quantized channels for all users K using Algorithm 1.
2: Schedule $2N_t$ users into N_t clusters using MPECG-SIR algorithm and compute $\boldsymbol{W}^{st}$ after strong users' selection.

Step 2) Power allocation & Superposition Coding

3: Find the optimal α_k for all clusters using (61).
4: Generate $2N_t$ symbols using QPSK modulation. Then, superimose them as,
5: **for all** $k = 1$ to N_t **do**

$$x_k = \sqrt{\alpha_k}s_k^{st} + \sqrt{1-\alpha_k}s_k^{we},$$

6: **end for**

Step 3) Precoding

7: Save the superimposed symbols:

$$\boldsymbol{x}_s = [x_1, \ldots, x_k, \ldots, x_{N_t}], \quad \boldsymbol{x}_s \in \mathbb{C}^{N_t\times 1}$$

8: Precode: $\boldsymbol{x}_s^{Pre} = \boldsymbol{W}^{st}\boldsymbol{x}_s, \quad \boldsymbol{x}_s^{Pre} \in \mathbb{C}^{N_t\times 1}$

Step 4) Limited CSI Sharing

9: Share the limited CSIs (CDIs and CQIs) between the users of each cluster k.
10: After limited CSI sharing, each user identifies whether it is strong or weak.

Step 5) Equalization & Symbol Detection with SIC Process

11: **for all** $k = 1$ to N_t **do**
12: Use SBM to equalize y_k^{st} as, $y_k^{\overline{st}} = \frac{\hat{\boldsymbol{h}}_k^{st}}{\boldsymbol{h}_k^{st}}y_k^{st}$.
13: Estimate $\hat{s}_k^{we_{st}}$ at the strong user side using (62).
14: Perform SIC and recover $\hat{s}_k^{st}$ at strong user using (63).
15: Use WBM to equalize y_k^{we} as, $y_k^{\overline{we}} = \frac{\hat{\boldsymbol{h}}_k^{st}}{\boldsymbol{h}_k^{we}}y_k^{we}$.
16: Recover $\hat{s}_k^{we_{we}}$ at weak user side using (64).
17: **end for**

IX. Simulation Results

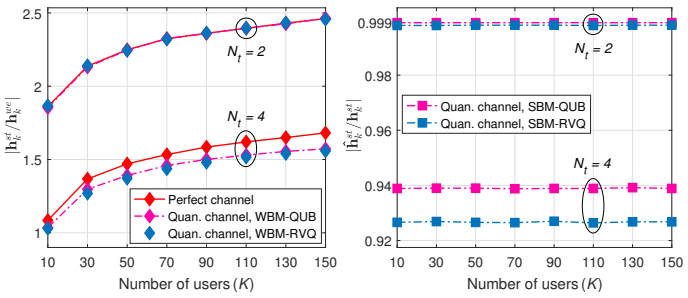
In this section, the performance of the proposed cooperative NOMA with WBM and SBM equalizers is evaluated in terms of sum-rate and BER. Furthermore, it is compared with other non-cooperative schemes such as, the adapative NOMA, and the TDMA [10] with both perfect CSIT and limited feedback cases. The simulation parameters are summarized in Table II.

TABLE II: SIMULATION PARAMETERS

Parameter	Value
Total power of each cluster, P	15 dB
SNR in BER test, ρ	$0,\ 4, \ldots,\ 28$ dB
Number of users, K	$10,\ 20, \ldots,\ 150$
Number of antennas of BS, N_t	$2, 4, \ldots,\ 10$
Number of clusters	N_t
Number of users per cluster	2
Number of supportable users	$2N_t$
Channel model	Rayleigh small scale fading
CSI feedback type	Perfect & limited feedback
Limited feedback model	RVQ & QUB
Number of feedback bits, B	6 & 8 bits/user
Modulation type	QPSK

A. Sum-Rate Performance

The achievable rates of strong and weak users are highly affected by the values of WBM and SBM equalizers. Therefore, before starting sum-rate tests, the average absolute ratios of these equalizers versus the number of users is shown in Fig. 3. The average absolute ratio of WBM equalizer $|\boldsymbol{h}_k^{st}/\boldsymbol{h}_k^{we}|$ is demonstrated in Fig. 3(a), which increases with the number of users K, since the channel quality of the strong user $\boldsymbol{h}_k^{st}$ is highly improved with K, due to increasing the multiuser diversity gain[4]. Thus, the absolute ratio $|\boldsymbol{h}_k^{st}/\boldsymbol{h}_k^{we}|$ increases with K. However, the value of $|\boldsymbol{h}_k^{st}/\boldsymbol{h}_k^{we}|$ decreases when N_t increases from 2 to 4, because when $N_t = 2$, the value of $\boldsymbol{h}_k^{st}$ over $\boldsymbol{h}_k^{we}$ is high and therefore, the absolute ratio $|\boldsymbol{h}_k^{st}/\boldsymbol{h}_k^{we}|$ is also high. Whereas, when N_t increases, the number of interference terms of each channel also increases and works to reduce the difference between $\boldsymbol{h}_k^{st}$ and $\boldsymbol{h}_k^{we}$. Thus, the absolute ratio $|\boldsymbol{h}_k^{st}/\boldsymbol{h}_k^{we}|$ decreases when N_t increases to 4.



(a) WBM equalizer (b) SBM equalizer

Fig. 3: The absolute ratio of equalization factors versus the number of users. $B = 8$ bits/user for the quantized channels.

Fig. 3(b) shows that the absolute ratio of SBM equalizer at strong users is not affected by K, and has lower values than that of WBM equalizer ($0.92 < |\hat{\boldsymbol{h}}_k^{st}/\boldsymbol{h}_k^{st}| < 1$), since the SBM absolute ratio is between similar channels, which are, the quantized channel of strong user and the perfect channel of strong user (i.e., $|\hat{\boldsymbol{h}}_k^{st}/\boldsymbol{h}_k^{st}|$). However, similar to WBM equalizer, the absolute ratio of SBM decreases when N_t increases. Also, the SBM-RVQ has a lower absolute ratio than the SBM-QUB, especially when $N_t = 4$, since the RVQ has relatively higher quantization errors than the QUB, which results in smaller $|\hat{\boldsymbol{h}}_k^{st}/\boldsymbol{h}_k^{st}|$ values than with QUB model.

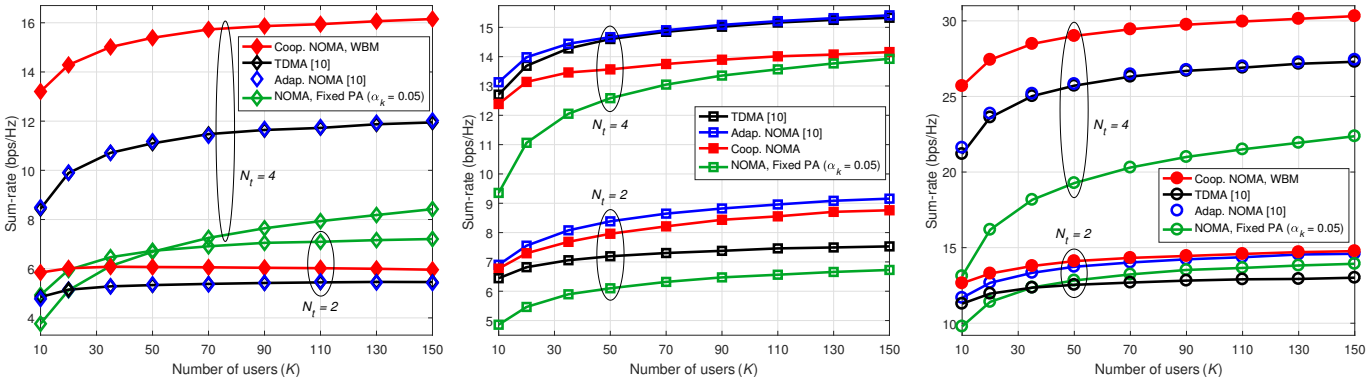
Note that to calculate the sum-rate with WBM or with SBM, the absolute ratio is squared and multiplied by the noise variance σ_n^2. Besides, with WBM, we have $|\boldsymbol{h}_k^{st}/\boldsymbol{h}_k^{we}| > 1$. Thus, the value of $|\boldsymbol{h}_k^{st}/\boldsymbol{h}_k^{we}|^2$ is amplified, (i.e., $|\boldsymbol{h}_k^{st}/\boldsymbol{h}_k^{we}|^2 > |\boldsymbol{h}_k^{st}/\boldsymbol{h}_k^{we}|$) and works to enhance the value of σ_n^2, and this limits the sum-rate of weak users, especially when $N_t = 2$. Contrarily, with SBM, we have $|\hat{\boldsymbol{h}}_k^{st}/\boldsymbol{h}_k^{st}| < 1$ for any N_t number, and hence, $|\hat{\boldsymbol{h}}_k^{st}/\boldsymbol{h}_k^{st}|^2 < |\hat{\boldsymbol{h}}_k^{st}/\boldsymbol{h}_k^{st}|$. Accordingly, the value of σ_n^2 is reduced with SBM equalizer and the sum-rate of strong users can be further enhanced when N_t increases in the case of limited feedback, as will be seen in the following results.

Fig. 4 shows the sum-rates of weak users, strong users, and the total system versus the number of users for the case of perfect CSIT. In Fig. 4(a), it can be seen that weak users with the proposed WBM equalizer in cooperative NOMA always outperforms those in adaptive NOMA and TDMA systems [10]. However, according to the PA constraint in the adaptive NOMA [10], the system is switched to the TDMA transmission in case of failure PA (which arises with a high ICI at weak user that causes $\alpha_k < 0$). Therefore, the actual NOMA performance with adaptive NOMA system cannot be determined. For this reason, we remove the adaptation from the adaptive NOMA system and test the actual NOMA performance with a fixed PA. Also, we allocate most of the total power to the weak user, i.e., $1 - \alpha_k = 0.95$, to see if it is possible to overcome the ICI by increasing the power. It can be seen that, when $N_t = 2$, we successfully increase the sum-rate of weak users (with a fixed PA) and make it more than those in other schemes. However, when $N_t = 4$, weak users in NOMA with fixed PA give the lowest sum-rate than in other schemes since the ICI is highly increased and causes a significant reduction in the sum-rate of weak users.

Fig. 4(a) also shows that the sum-rate of weak users with WBM equalizer achieves the best sum-rate when $N_t = 4$, since the equalization factor of WBM is highly decreased when $N_t = 4$, and consequently, better data rate can be obtained for each weak user after equalization.

In Fig. 4(b), it is observed that the sum-rate of strong users in adaptive NOMA is higher than in TDMA system when $N_t = 2$. whether, when $N_t = 4$, strong users in adaptive NOMA have almost the same sum-rate as in TDMA system, because the adaptive NOMA is switched to the TDMA transmission due to increasing ICI at weak users (as mentioned earlier). However, strong users in actual NOMA system represented by the fixed PA scheme have the lowest sum-rate performance due to allocating a small power portion for each strong user (i.e., $\alpha_k = 0.05$) and assigning most of the total power for weak users (i.e., $1 - \alpha_k = 0.95$), in order to raise weak users' sum-rate. On the other hand, strong users in the proposed cooperative NOMA achieve higher sum-rate than in TDMA when $N_t = 2$, while they have lower sum-rate than in TDMA when $N_t = 4$, because our PA strategy aims to achieve more balanced performance for both NOMA users.

[4]The channel of weak user is not highly affected by the multiuser diversity gain, since weak user selection is not only based on a high channel quality as weak user's channel should also be correlated to the channel of strong user in order to have a better use of a single BF vector for both users.



(a) Weak users (b) Strong users (c) Total system

Fig. 4: Sum rate versus the number of users with perfect CSIT.

The total sum-rate is shown in Fig. 4(c). It can be noticed that the proposed cooperative NOMA achieves the best performance over other schemes when $N_t = 2$ and 4. Also, when $N_t = 4$, our proposed system with WBM equalizer can achieve a much higher sum-rate than other schemes due to the high reduction in the SINR of each weak user after decreasing the values of WBM equalizer when $N_t = 4$ (as shown in Fig. 3(a)). Also, when $N_t = 4$, it can be seen that, the adaptive NOMA has compatible performance with TDMA, which indicates that the adaptive NOMA is switched to the TDMA transmission. Finally, the sum-rate of NOMA with fixed PA has the lowest performance due to the high ICI at weak users and wasting most of the total power to raise up weak users' sum-rate.

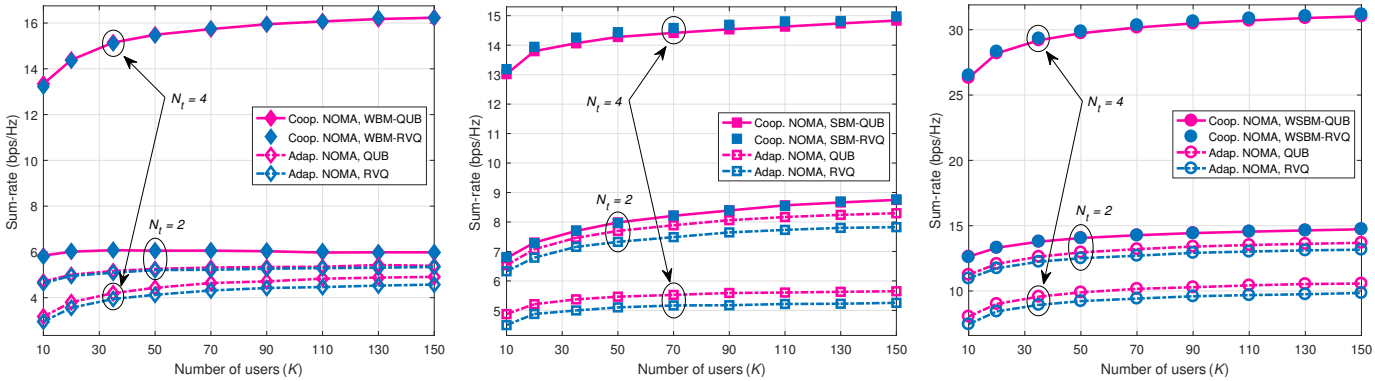
Fig. 5 shows the achievable rates of NOMA users and the total system in the case of limited feedback. Note that the SBM equalizer is activated here at strong users to eliminate the ICI caused by quantization errors. The total system sum-rate with cooperative NOMA is denoted by "WSBM" in Fig. 5(c) to indicate that both WBM and SBM are used.

The proposed cooperative NOMA is compared in Fig. 5 with the adaptive NOMA system based on two feedback models, the QUB model, and the conventional RVQ model. It is observed that weak users, strong users, and the total system can achieve the best performance with our proposed system based on QUB and RVQ. Besides, strong users with RVQ achieve a slightly higher sum-rate than with QUB, especially, when, $N_t = 4$ due to SBM equalization effect (refer to Fig. 3(b)) that can reduce the noise variance and leads to a slight increment in the sum-rate of strong users and the total system. Moreover, it can be seen in Fig. 5 that with cooperative NOMA, the sum-rates of weak users, strong users, and the total system are highly improved when N_t increases to 4, because with perfect ICI cancellation using the proposed equalizers, increasing N_t adds more users and therefore the sum-rates of NOMA users and the total system become higher. Contrarily, with adaptive NOMA, the sum-rates of NOMA users become worse when N_t increases, since increasing N_t adds more ICI in the case of limited feedback and consequently, the SINR is decreased at each user in the adaptive NOMA. Therefore, unlike the proposed cooperative NOMA, increasing N_t in the adaptive NOMA deteriorates the sum-rate performance.

Furthermore, with adaptive NOMA, the weak users, strong users, and the total system have better performance with QUB than with RVQ when $N_t = 2$ and 4, since the QUB has relatively smaller quantization errors than RVQ. However, in Fig. 5(a) when $N_t = 2$, the sum-rate of weak users with QUB is almost the same as with RVQ in adaptive NOMA, while in Fig. 5(b), the sum-rate of strong users with QUB is better than with RVQ. The reason for these different performances is related to the PA strategy in [10]. To be more specific, the PA is firstly determined for weak user, and when the SINR of the weak user is dropped due to the quantization error, the BS compensates this decrement by allocating more power to the weak user and this power is taken from the power portion of the strong user. Therefore, the effect of quantization error mostly appears at the sum-rate of strong users in adaptive NOMA, which in fact includes the effects of quantization errors of both weak and strong users. However, because the QUB has relatively smaller quantization errors than the RVQ, the sum-rate of strong users is better with QUB than with RVQ. On the other hand, when $N_t = 4$, the adaptive NOMA is switched to the TDMA transmission due to increasing ICI at weak users. Therefore, the effect of quantization error appears at the sum-rate performances of both users.

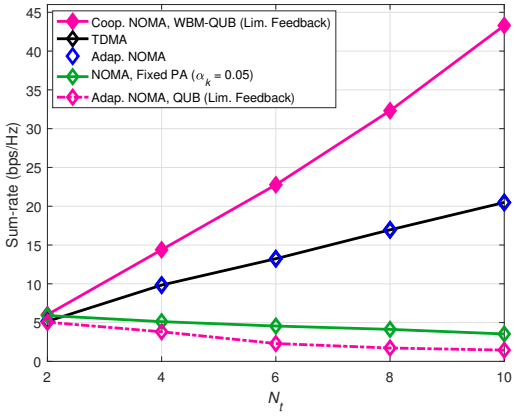
Fig. 6 illustrates the sum-rates of weak users and the total system versus the number of transmit antennas N_t. The sum-rate of strong users is combined within the total sum-rate. In Fig. 6(a), it can be seen that the sum-rate of weak users in NOMA with fixed PA decreases with N_t, because more ICI is added when N_t increases. This degraded performance also limits the total system sum-rate as shown in Fig. 6(b). Besides, weak users and the total system have the same performance in adaptive NOMA and TDMA system, which indicates that NOMA is switched to the TDMA transmission due to increasing ICI at weak users.

In the case of limited feedback, it can be noticed in Figs. 6(a) and 6(b) that the sum-rates of weak users and the total system decrease with N_t in adaptive NOMA, because the ICI occurs at both strong and weak users and becomes higher when N_t increases. Contrarily, with the proposed cooperative NOMA, the sum-rates of weak users and the total system increase linearly with N_t and achieve the best performance over other schemes, as the ICI can be totally removed at strong and weak users with the use of the proposed equalizers.

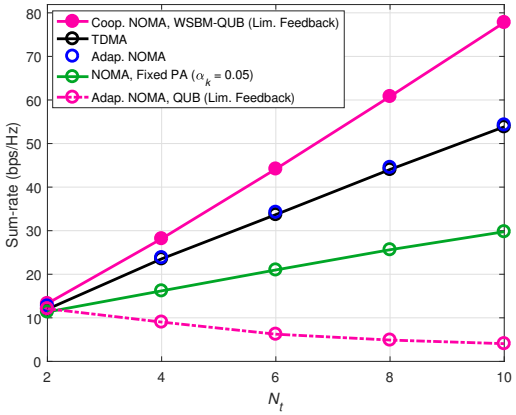


(a) Weak users (b) Strong users (c) Total system

Fig. 5: Sum-rate versus the number of users with limited feedback. $B = 8$ bits/user.

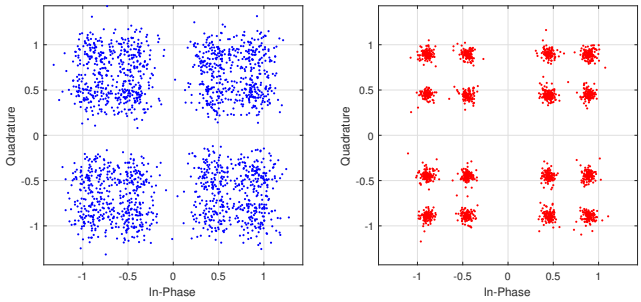


(a) Weak users



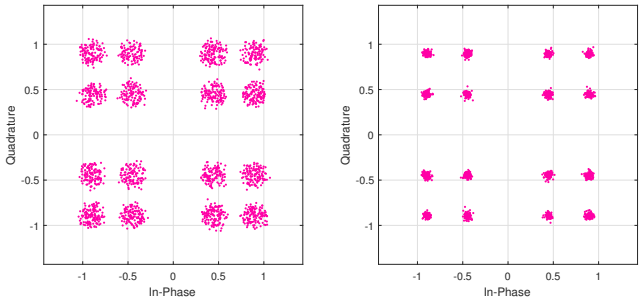
(b) Total system

Fig. 6: Sum-rate versus the number of transmit antennas N_t. $K = 20$ users, and $B = 8$ bits/user for limited feedback cases.



(a) Without WBM (Adap. NOMA) (b) With WBM

Fig. 7: Constellation of the superimposed signals with QPSK modulation at weak users with perfect CSIT. $N_t = 2$, $K = 40$ users, $\alpha_k = 0.1$, and $\rho = 24$ dB.



(a) Without SBM (b) With SBM

Fig. 8: Constellation of the superimposed signals with QPSK at strong users with limited feedback (QUB model). $N_t = 2$, $K = 40$ users, $\alpha_k = 0.1$, $\rho = 24$ dB, and $B = 6$ bits/user.

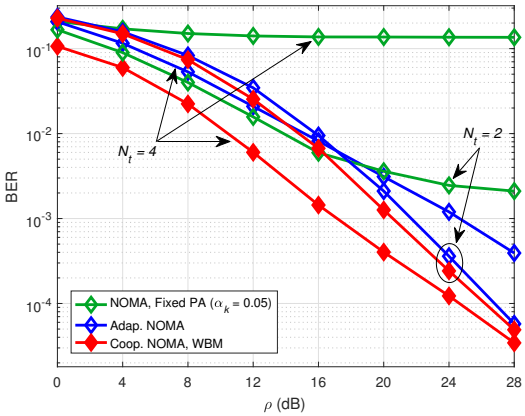
B. Bit Error Rate (BER) Performance

In this subsection, the BER performance of NOMA users with QPSK modulation is tested with the proposed cooperative NOMA and compared with those in non-cooperative schemes. In this test, we set $K = 20$ users to examine the robustness of the proposed system with a relatively low multiuser diversity. However, before starting the BER test, Figs. 7 and 8 show the effect of using the proposed equalizers on the constellation of the received signals at NOMA users. Note that a specific α_k is chosen in these Figs. to focus on the equalization effect.

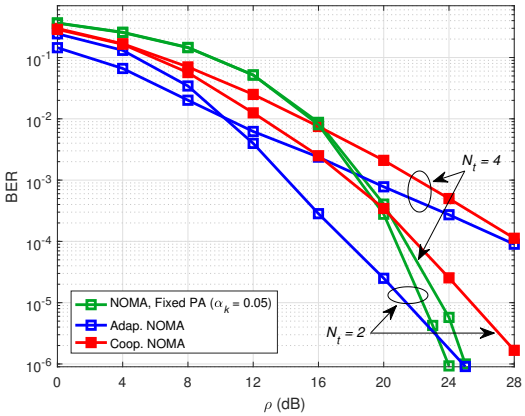
Fig. 7 depicts the superimposed signal constellations of weak users with and without using WBM equalizer. It is noteworthy that even with perfect CSIT, the symbols are not concentrated properly in Fig. 7(a), because of the high ICI. Fig. 7(b), shows the effect of using WBM on correcting the constellation, which leads to better BER performance.

Fig. 8 illustrates how the use of SBM equalizer can correct the constellation of the received signals at strong users in the case of limited feedback. By comparing the effects of using the proposed equalizers in Figs. 7(b) and 8(b), it can be observed that a better constellation is obtained after using SBM equalizer than the obtained after using WBM equalizer, because with WBM the noise is highly amplified when $N_t = 2$

(as stated previously). Also, by comparing Figs. 7(a) and 8(a), it can be noticed that, the resulted ICI from BF and channel mismatch in the adaptive NOMA has a worse effect than the generated ICI in the case of limited feedback.



(a) Weak users



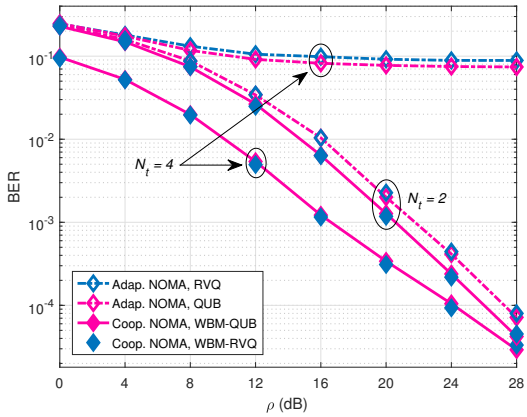
(b) Strong users

Fig. 9: BER versus SNR (ρ) with perfect CSIT. $K = 20$ users.

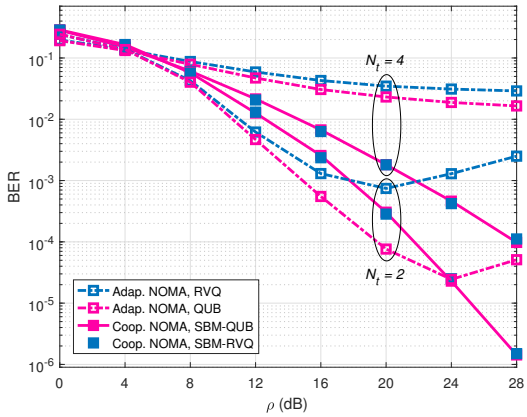
In Fig. 9, the BER performance of weak and strong users is tested in the case of perfect CSIT. Fig. 9(a) shows that the best performance of weak users can be achieved with WBM equalizer in cooperative NOMA when $N_t = 2$ and 4, which is a similar scenario to that in the sum-rate test. Also, when $N_t = 2$, the weak users in NOMA with fixed PA have the lowest BER performance even when allocating most of the total power to the weak users, i.e., $1 - \alpha_k = 0.95$, because of the generated ICI at weak users due to BF and channel mismatch. When $N_t = 4$, the weak users performance in NOMA with fixed PA becomes even worse, because of increasing ICI at the receiver of each weak user after increasing N_t.

On the other hand, strong users in Fig. 9(b) have better BER performance in NOMA with fixed PA even with a small PA portion ($\alpha_k = 0.05$), because of perfect matching between strong users' channels and the designed beams, which leads to a perfect ICI cancellation and hence better BER performance. However, when $N_t = 2$, the strong users in NOMA with fixed PA have better BER performance than when $N_t = 4$, because the multiuser diversity gain increases with the reduction of N_t.

Fig. 9 also shows that, when $N_t = 4$, the BER performance of strong users in adaptive NOMA is better than the BER of these users in cooperative NOMA. This is due to the difference in PA strategies, wherein, in the adaptive NOMA, more power is allocated to strong users, while in cooperative NOMA, the proposed PA offers more balanced performance for both users.



(a) Weak users



(b) Strong users

Fig. 10: BER versus SNR (ρ) with limited feedback. $B = 8$ bits/user and $K = 20$ users.

In Fig. 10, the BER performance of weak and strong users is tested with limited feedback. We observe in Fig. 10(a) that the BER performance of weak users with QUB is the same as with RVQ in the adaptive NOMA system, while in Fig. 10(b), the BER performance of strong users with QUB is better than with RVQ. This is a similar performance to the sum-rate of weak and strong users in Figs. 5(a) and 5(b) and as stated previously, it is related to the PA strategy in [10]. Accordingly, the effect of limited feedback mostly appears at the BER performance of strong users in the adaptive NOMA. Also, it is observed that, when $N_t = 2$ and $\rho > 20$, the BER performance of strong users with RVQ becomes worse. This performance is related to the ICI, i.e., the ICI term $\sqrt{\rho}\sum_{j\neq k} \boldsymbol{h}_k^{we}\boldsymbol{w}_j^{st}$ is increased with increasing ρ value. Therefore, the power increment in this case degrades the BER performance. A similar performance appears with QUB when $\rho > 24$, as the QUB has relatively smaller quantization errors than the RVQ. However, as was seen in the sum-rate tests, when $N_t = 4$, the adaptive NOMA is switched to the TDMA transmission due to increasing ICI at weak users. Thus, the effect of limited feedback appears at the BER performance of both users as NOMA PA strategy is not

applied to TDMA system. On the other hand, in cooperative NOMA, when $N_t = 2$ and 4, both NOMA users have the same performance with QUB and RVQ, since the SINRs of weak users do not decrease with the use of WBM equalizer.

X. Conclusion

This paper considered the problem of imperfect ICI cancellation at weak users in MU-MISO-NOMA system, due to precoding weak users' signals by the BF vectors of strong users. To solve this problem, a new cooperative NOMA was introduced based on D2D CSI sharing between the clustered strong and weak users. Relying on this cooperation, the WBM equalizer was proposed at the receivers of weak users in order to remove the generated ICI in the cases of perfect CSIT and limited feedback. It has been shown that the D2D CSI sharing is also required for each user to determine whether it is strong or weak to decide accordingly whether to perform the SIC process or not.

In addition, the SBM equalizer was proposed at the receivers of strong users to cancel the generated ICI in the case of limited feedback. Furthermore, a new PA strategy was proposed, which can enhance the sum-rate of weak users over that in [10], by considering the gained throughput from interference cancellation at weak users. Finally, the BER of NOMA users with the proposed WBM and SBM equalizers has been tested and compared with those in non-cooperative NOMA schemes.

Simulation results clarified that in the case of perfect CSIT, increasing N_t in non-cooperative NOMA schemes either degrades the performance or switches the system to TDMA transmission mode due to increasing ICI at weak users. Whereas in the proposed cooperative NOMA, the ICI can be suppressed effectively, and hence increasing N_t leads to better performance. In the case of limited feedback, the adaptive NOMA has better performance with the QUB model than with conventional RVQ. On the other hand, the cooperative NOMA has approximately similar performance with these models. Finally, the results showed that the proposed cooperative NOMA system with WBM and SBM equalizers significantly outperforms other schemes in terms of sum-rate and BER while maintaining comparable performance between strong and weak users with both feedback scenarios.

References

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5g be?" *IEEE Journal on selected areas in communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[2] G. T. 38.913, "Study on scenarios and requirements for next generation access technologies," 2016.

[3] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6g: Ai empowered wireless networks," *IEEE Communications Magazine*, vol. 57, no. 8, pp. 84–90, 2019.

[4] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiple access for 5g: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, 2015.

[5] Z. Wei, J. Yuan, D. W. K. Ng, M. Elkashlan, and Z. Ding, "A survey of downlink non-orthogonal multiple access for 5g wireless communication networks," *arXiv preprint arXiv:1609.01856*, 2016.

[6] S. Norouzi, A. Morsali, and B. Champagne, "Optimizing transmission rate in noma via block diagonalization beamforming and power allocation," in *2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*. IEEE, 2019, pp. 1–5.

[7] J. Zhu, J. Wang, Y. Huang, F. Fang, K. Navaie, and Z. Ding, "Resource allocation for hybrid noma mec offloading," *IEEE Transactions on Wireless Communications*, 2020.

[8] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2744–2757, 2017.

[9] B. Kimy, S. Lim, H. Kim, S. Suh, J. Kwun, S. Choi, C. Lee, S. Lee, and D. Hong, "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *MILCOM 2013-2013 IEEE Military Communications Conference*. IEEE, 2013, pp. 1278–1283.

[10] S. Liu, C. Zhang, and G. Lyu, "User selection and power schedule for downlink non-orthogonal multiple access (noma) system," in *2015 IEEE International Conference on Communication Workshop (ICCW)*. IEEE, 2015, pp. 2561–2565.

[11] M. M. Al-Wani, A. Sali, B. M. Ali, A. A. Salah, K. Navaie, C. Y. Leow, N. K. Noordin, and S. Hashim, "On short term fairness and throughput of user clustering for downlink non-orthogonal multiple access system," in *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*. IEEE, 2019, pp. 1–6.

[12] M. M. Al-Wani, A. Sali, N. K. Noordin, S. J. Hashim, C. Y. Leow, and I. Krikidis, "Robust beamforming and user clustering for guaranteed fairness in downlink noma with partial feedback," *IEEE Access*, vol. 7, pp. 121 599–121 611, 2019.

[13] Y. Fu, M. Zhang, L. Salaun, C. W. Sung, and C. S. Chen, "Zero-forcing oriented power minimization for multi-cell miso-noma systems: A joint user grouping, beamforming and power control perspective," *IEEE Journal on Selected Areas in Communications*, 2020.

[14] T. H. Kim, R. W. Heath, and S. Choi, "Multiuser mimo downlink with limited feedback using transmit-beam matching," in *2008 IEEE International Conference on Communications*. IEEE, 2008, pp. 3506–3510.

[15] J. Song, B. Lee, S. Noh, and J.-H. Lee, "Limited feedback designs for machine-type communications exploiting user cooperation," *IEEE Access*, vol. 7, pp. 95 154–95 169, 2019.

[16] J. Chen, H. Yin, L. Cottatellucci, and D. Gesbert, "Feedback mechanisms for fdd massive mimo with d2d-based limited csi sharing," *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 5162–5175, 2017.

[17] M. Choi, D.-J. Han, and J. Moon, "Bi-directional cooperative noma without full csit," *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7515–7527, 2018.

[18] S. Zhou, Z. Wang, and G. B. Giannakis, "Quantifying the power loss when transmit beamforming relies on finite-rate feedback," *IEEE Transactions on Wireless Communications*, vol. 4, no. 4, pp. 1948–1957, 2005.

[19] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-antenna downlink channels with limited feedback and user selection," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1478–1491, 2007.

[20] J. Chen, H. Yin, L. Cottatellucci, and D. Gesbert, "Dual-regularized feedback and precoding for d2d-assisted mimo systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 10, pp. 6854–6867, 2017.

[21] D. J. Love and R. W. Heath, "Limited feedback unitary precoding for spatial multiplexing systems," *IEEE Transactions on Information theory*, vol. 51, no. 8, pp. 2967–2976, 2005.

[22] N. Jindal, "Mimo broadcast channels with finite-rate feedback," *IEEE Transactions on information theory*, vol. 52, no. 11, pp. 5045–5060, 2006.

[23] C. K. Au-Yeung and D. J. Love, "On the performance of random vector quantization limited feedback beamforming in a miso system," *IEEE Transactions on Wireless Communications*, vol. 6, no. 2, pp. 458–462, 2007.

[24] A. K. Gupta and S. Nadarajah, *Handbook of beta distribution and its applications*. CRC press, 2004.

[25] K. K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On beamforming with finite rate feedback in multiple-antenna systems," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2562–2579, 2003.

[26] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on selected areas in communications*, vol. 24, no. 3, pp. 528–541, 2006.

[27] Y. Shi, Q. Yu, W. Meng, and Z. Zhang, "Maximum product of effective channel gains: an innovative user selection algorithm for downlink multi-user multiple input and multiple output," *Wireless Communications and Mobile Computing*, vol. 14, no. 18, pp. 1732–1740, 2014.

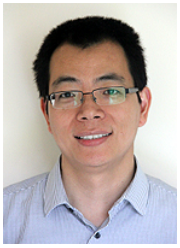
[28] J. Cui, Z. Ding, and P. Fan, "A novel power allocation scheme under outage constraints in noma systems," *IEEE Signal Processing Letters*, vol. 23, no. 9, pp. 1226–1230, 2016.

Mohanad M. Al-Wani received the B.Sc. degree in electrical engineering from Al-Mustansiriya University, Baghdad, Iraq, in 2006, and the M.Sc. degree in wireless communication engineering from University of Technology, Baghdad, Iraq, in 2010. He is currently pursuing the Ph.D. degree with Universiti Putra Malaysia. From 2010 to 2011, he worked as a Lecturer with the Department of Computer Science, Dijlah University College, Baghdad, Iraq. From 2019 to 2020, he was a visiting researcher at the KIOS Research Center, University of Cyprus, Cyprus. His research interests include new multiple access technologies, beamforming, resource scheduling, limited feedback techniques, cooperative wireless networks, and enabling technologies for 5G and beyond.

Aduwati Sali Professor Ir. Dr. Aduwati Sali is currently a Professor at Department of Computer and Communication Systems, Faculty of Engineering, Universiti Putra Malaysia (UPM) since February 2019. She was a Deputy Director at UPM Research Management Centre (RMC) responsible for Research Planning and Knowledge Management from 2016 to 2019. She obtained her Ph.D. in Mobile and Satellite Communications form University of Surrey, UK, in July 2009; MSc in Communications and Network Engineering from UPM, Malaysia, in April 2002 and BEng in Electrical Electronics Engineering (Communications) from University of Edinburgh, UK, in 1999. She is also a Chartered Engineer (C. Eng) registered under UK Engineering Council and a Professional Engineer (P. Eng.) under Board of Engineers Malaysia (BEM). She worked as an Assistant Manager with Telekom Malaysia Bhd from 1999 until 2000. She is involved with IEEE as a Chair to ComSoc/VTS Malaysia (2017 and 2018) and Young Professionals (YP) (2015); Young Scientists Network-Academy of Sciences Malaysia (YSN-ASM) as an Honorary Member (since 2020), Chair (2018) and Co-Chair (2017) for Science Policy. She is also the recipient of 2018 Top Research Scientists Malaysia (TRSM) Award from Academy of Sciences Malaysia (ASM).

Sumaya D. Awad received the B.Sc. degree in electronic and communications engineering and the M.Sc. degree in electronic engineering from the Electronic and Communications Engineering Department, University of Technology, Baghdad, Iraq, in 2007 and 2010, respectively. She is currently pursuing the Ph.D. degree with Universiti Putra Malaysia. From 2010 to 2016, she worked as a Lecturer with the Communications Engineering Department, Al-Ma'moon University College, Iraq. From 2019 to 2020, she was a visiting researcher at the KIOS Research Center, University of Cyprus, Cyprus. Her research interests include multibeam satellite communication, beamforming, multiuser diversity techniques, and fade mitigation techniques.

Asem A. Salah Dr. Asem A. Salah received the PhD and MSc degrees in Communications and Network Engineering from Universiti Putra Malaysia (UPM), Malaysia, in May 2015, and March 2011, respectively, and the BSc degree in Telecommunication Technology from the Arab American University-Jenin (AAUJ), Palestine, in June 2005. He is currently at the department of Computer System Engineering, Faculty of Engineering and Information Technology, Arab American University, Palestine. He worked for the Department of Electrical Engineering, Faculty of Engineering, University of Malaya from Dec. 2018 to Dec. 2019. Before that he worked for the Department of Computer and Communication Systems Engineering, Faculty of Engineering, UPM from 2015 to 2018. During his work he has been involved in several research projects, in Wireless Communications, Mobile Communication Systems, 5G, Wireless Sensors Network, D2D Communications, Deep Learning Algorithms, Passive Radar, and FSR.



Zhiguo Ding (Fellow, IEEE) received his B.Eng in Electrical Engineering from the Beijing University of Posts and Telecommunications in 2000, and the Ph.D degree in Electrical Engineering from Imperial College London in 2005. From Jul. 2005 to Apr. 2018, he was working in Queen's University Belfast, Imperial College, Newcastle University and Lancaster University. Since Apr. 2018, he has been with the University of Manchester as a Professor in Communications. From Oct. 2012 to Sept. 2020, he has also been an academic visitor in Princeton University. Dr Ding' research interests are 5G networks, game theory, cooperative and energy harvesting networks and statistical signal processing. He is serving as an Area Editor for the IEEE Open Journal of the Communications Society, an Editor for IEEE Transactions on Communications, IEEE Transactions on Vehicular Technology, and Journal of Wireless Communications and Mobile Computing, and was an Editor for IEEE Wireless Communication Letters, IEEE Communication Letters from 2013 to 2016. He received the best paper award in IET ICWMC-2009 and IEEE WCSP-2014, the EU Marie Curie Fellowship 2012-2014, the Top IEEE TVT Editor 2017, IEEE Heinrich Hertz Award 2018, IEEE Jack Neubauer Memorial Award 2018, IEEE Best Signal Processing Letter Award 2018, and Web of Science Highly Cited Researcher 2019.

Nor K. Noordin Nor Kamariah Noordin graduated from University of Alabama in 1987 and obtained her Ph.D. from Universiti Putra Malaysia in the area of Wireless and Communication Engineering. She is currently the Dean of Engineering of UPM. Apart from wireless her research interest also includes education engineering. She has published more than 300 journals, book chapters and conference papers. She has led more than 20 research projects, funded by local and international grant providers.

Shaiful J. Hashim received the B.Eng. degree from the University of Birmingham, U.K., in 1998, the M.Sc. degree from the National University of Malaysia, in 2003, and the Ph.D. degree from Cardiff University, U.K., in 2011, all in electrical and electronics engineering. He is currently an Associate Professor with the Department of Computer and Communication Systems Engineering, Faculty of Engineering, Universiti Putra Malaysia (UPM). He has contributed to more than 100 technical and research publications. His research interests include cloud computing, the Internet of Things (IoT), network security, and nonlinear wireless measurement systems. He is one of the winners of the prestigious IEEE MTT-11 2008 Creativity and Originality in Microwave Measurements Competition.

Chee Yen (Bruce) Leow received the B.Eng. degree in computer engineering from Universiti Teknologi Malaysia (UTM), Johor, Malaysia, and the Ph.D. degree from Imperial College London, London, U.K., in 2007 and 2011, respectively. Since 2007, he has been an Academic Staff with the Faculty of Electrical Engineering, UTM. He is currently an Associate Professor with the Faculty and a Research Fellow with the Wireless Communication Centre, Higher Institution Centre of Excellence, and UTM-Ericsson Innovation Centre for 5G. His research interests include cooperative communication, MIMO, drone communication, physical layer security and 5G.