ORIGINAL PAPER

Multi-cluster MIMO non-orthogonal multiple access for multi-cell systems

Changyong Shin[1] 🆔

Abstract

In this paper, we present a multi-cluster multiple-input multiple-output non-orthogonal multiple access technique to improve spectral efficiency by simultaneously transmitting multiple data streams to multiple clusters in multi-cell systems. This technique employs receive beamforming matrices for cell-edge users that minimize the multi-cell interference power, irrespective of the number of interfering cells. Based on the receive beamforming matrices, the technique designs transmit beamforming matrices for a serving base station (BS) that eliminate inter-cluster interference (ICI) between cell-edge users and maximize the lower bounds of the received SNRs at the cell-edge users. Then, the technique uses the transmit beamforming matrices to find receive beamforming matrices for cell-center users that eliminate ICI between the cell-center users and maximize the lower bounds of the received SNRs at the cell-center users. This technique does not require channel feedback from the users to the serving BS in order to find all the receive beamforming matrices, thereby reducing system overhead. In addition, based on the transmit and receive beamforming matrices, the proposed technique allocates the transmit power to clusters and users of each cluster in a manner that maximizes the sum rate performance under power constraints supporting fairness between clusters and between users in each cluster. We demonstrate through simulations that the proposed technique achieves better sum rate and outage performances than the existing techniques and can provide a good balance between the performances and the number of clusters simultaneously supported by the serving BS.

Keywords Non-orthogonal multiple access (NOMA) · Multi-cluster multiple-input multiple-output (MC-MIMO) · Multi-cell system · Spectral efficiency · Interference

1 Introduction

Wireless data traffic is explosively increasing due to a rapidly growing number of wireless access devices and the increase in demand for various high-quality wireless services [1]. In order to cope with this soaring wireless data traffic, next-generation wireless communication systems are actively discussed [2–5]. Since those systems should offer high transmission capacity with limited spectrum resources, techniques improving spectral efficiency receive a lot of attention [6–12]. Specifically, non-orthogonal multiple access (NOMA) may provide higher spectral efficiency than orthogonal multiple access, which has been widely employed in existing wireless communication systems, by simultaneously serving multiple users utilizing the same time and frequency resource [13, 14]. Therefore, NOMA is considered a promising technique for next-generation wireless communication systems [15, 16].

To enhance spectral efficiency in single-input single-output (SISO) and multiple-input multiple-output (MIMO) cellular systems, the NOMA techniques using successive interference cancellation (SIC) are proposed in [17]. In the MIMO-NOMA scheme in [18], a base station (BS) employs random beamforming and intra-cluster superposition coding and a receiver exploits SIC. Although the techniques decrease system overhead, they cannot remove inter-cluster interference (ICI) and cannot transmit multiple data streams to each cluster. In addition, the techniques treat multi-cell interference (MCI) as noise and do not attempt to mitigate it. The multiple-input single-output NOMA technique in [19] combines zero-forcing

✉ Changyong Shin
 cyshin@sunmoon.ac.kr

[1] Department of Smart Information and Communications Engineering, Sun Moon University, Asan, South Korea

beamforming (ZFBF) with the hybrid NOMA precoding in [20] to minimize inter- and intra-cluster interference. However, this technique cannot be extended to MIMO systems straightforwardly and does not consider MCI.

The MIMO-NOMA technique in [21] provides low-complexity beamforming using singular value decomposition based on multiple-user channel state information (CSI). Although this technique is applicable for NOMA transmission to a single cluster with multiple users, it cannot be used to simultaneously transmit data streams to multiple clusters. To cancel ICI, the NOMA scheme in [22] employs ZFBF at a transmitter in a system with a single antenna at each receiver. However, this scheme requires a large number of transmit antennas for ICI cancellation and cannot be simply extended to a system with multiple antennas at each receiver. To eliminate ICI by using a small number of transmit antennas in a system with multiple antennas at each receiver, the MIMO-NOMA technique in [23] utilizes receive beamforming vectors orthogonal to ICI at each receiver. The MIMO-NOMA technique in [24] uses signal alignment to decrease the number of receive antennas necessary for ICI elimination.

However, the techniques in [21–24] cannot transmit multiple data streams to each cluster and do not deal with interference from adjacent cells. To simultaneously eliminate ICI and interference from one adjacent cell experienced by cell-edge users, the MIMO-NOMA techniques in [25] utilize interference alignment to design coordinated beamforming vectors. The techniques employ a practical user clustering where NOMA may provide a significant performance gain. However, they cannot transmit multiple data streams to each cluster and cannot deal with interference from more than one adjacent cell. To transmit multiple data streams to a cluster, the MIMO-NOMA technique in [26, 27] exploits the generalized singular value decomposition (GSVD) [28] to convert a MIMO-NOMA system into multiple SISO-NOMA systems. The technique in [29] uses the layered transmission schemes in [30, 31] for MIMO-NOMA transmissions and effectively allocates the transmit power to data streams for the layered transmissions.

Although the techniques in [26, 27, 29] may transmit multiple data streams to a single cluster, they are unable to transmit multiple data streams to multiple clusters simultaneously. Furthermore, they cannot handle interference from adjacent cells. The MIMO-NOMA technique in [32] considers a two-cell MIMO-NOMA interference channel and designs transmit beamforming matrices that maximize a weighted sum rate of two users, each with the largest channel gain in its own cell. However, this technique cannot support simultaneous transmission to multiple clusters in a cell and cannot be extended to deal with interference from more than one adjacent cell.

To improve spectral efficiency by simultaneously transmitting multiple data streams to multiple clusters in multi-cell systems, we propose a multi-cluster (MC) MIMO-NOMA beamforming technique and a power allocation scheme for beamformed data streams. This beamforming technique employs receive beamforming matrices for cell-edge users that minimize the MCI power, irrespective of the number of interfering cells. By utilizing the receive beamforming matrices, the technique finds transmit beamforming matrices for a serving BS that eliminate ICI between cell-edge users and maximize the lower bounds of the received SNRs at the cell-edge users. Then, the technique uses the transmit beamforming matrices to design receive beamforming matrices for cell-center users that eliminate ICI between the cell-center users and maximize the lower bounds of the received SNRs at the cell-center users.

Since this technique does not require channel feedback from the users to the serving BS in order to design all the receive beamforming matrices, it reduces system overhead. Furthermore, based on the transmit and receive beamforming matrices, the proposed scheme allocates the transmit power to clusters and users of each cluster in a manner that maximizes the sum rate performance under power constraints supporting fairness between clusters and between users in each cluster. It is demonstrated through simulations that the proposed MC-MIMO-NOMA method obtains better sum rate and outage performances than the existing techniques and can offer a good balance between the performances and the number of clusters simultaneously supported by the serving BS.

We note that the proposed MIMO-NOMA technique is distinguished from the existing techniques in the following points:

- It can transmit multiple data streams to multiple clusters simultaneously.
- It deals with interference from multiple neighboring cells and effectively mitigates the interference.
- It is applicable regardless of the number of interfering cells.
- It offers a trade-off between the number of clusters to serve simultaneously and the number of data streams to be transmitted to each cluster, thereby providing a good balance between the sum rate and outage performances and the number of clusters served simultaneously.
- It allocates the transmit power to clusters and users of each cluster to maximize the sum rate performance under power constraints that support fairness between clusters and between users in each cluster.

The rest of the paper is organized as follows. In Sect. 2, we describe the MC-MIMO-NOMA system model experiencing MCI. In Sect. 3, we state the proposed transmit and

Table 1 List of abbreviations

AWGN	Additive white Gaussian noise
BS	Base station
CSI	Channel state information
GSVD	Generalized singular value decomposition
ICI	Inter-cluster interference
MCI	Multi-cell interference
MC-MIMO	Multi-cluster multiple-input multiple-output
MC-MIMO-NOMA	Multi-cluster multiple-input multiple-output non-orthogonal multiple access
MIMO	Multiple-input multiple-output
MIMO-NOMA	Multiple-input multiple-output non-orthogonal multiple access
NOMA	Non-orthogonal multiple access
SC-MIMO-NOMA	Single cluster multiple-input multiple-output non-orthogonal multiple access
SIC	Successive interference cancellation
SISO	Single-input single-output
SISO-NOMA	Single-input single-output non-orthogonal multiple access
SNR	Signal-to-noise ratio
SVD	Singular value decomposition
ZFBF	Zero-forcing beamforming

receive beamforming technique for the MC-MIMO-NOMA system and the proposed power allocation scheme for beamformed data streams. Section 4 provides the simulation results demonstrating the sum rate and outage performances of the proposed MC-MIMO-NOMA method. Finally, some concluding remarks are presented in Sect. 5. The notations used in this paper are as follows. Matrices and vectors are denoted by symbols in boldface. $\text{tr}\{\cdot\}$ indicates the trace of a square matrix. $\det(\cdot)$ means the determinant of a square matrix. $\text{diag}\{\cdot\}$ represents the diagonal matrix with the given elements as entries in the main diagonal. $\mathbf{A}(:,i)$ and $\mathbf{A}(i,i)$ mean the ith column and the ith main diagonal element of the matrix $\mathbf{A}$, respectively. $\mathbf{I}_m$ and $\mathbf{0}$ represent the identity matrix with dimension m and a zero matrix or vector of appropriate dimensions, respectively. $(\cdot)^H$, $E\{\cdot\}$, and $\|\cdot\|$ represent Hermitian, statistical expectation, and l_2-norm, respectively. Table 1 details the abbreviations used in this paper.

2 Multi-cell MC-MIMO-NOMA system model

In this paper, we consider a MC-MIMO-NOMA system experiencing MCI from L adjacent cells as shown in Fig. 1. The serving cell BS simultaneously transmits data streams to K clusters with M transmit antennas. Two users, which are a cell-center user and a cell-edge user, exist in each cluster and NOMA is applied for data transmission to the users. Since
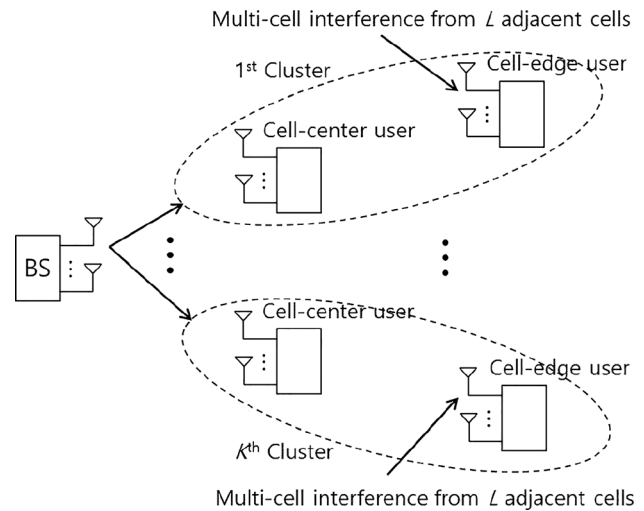

Fig. 1 MC-MIMO-NOMA system suffering from multi-cell interference

this user pairing for NOMA provides high spectral efficiency [14, 17], the user pairing is widely adopted in NOMA systems. Each user receives $J\,(\leq N)$ data streams with N receive antennas. In addition, the BSs of the L interfering cells are also equipped with M transmit antennas.

Denoting the channel from the serving cell BS to the cell-edge user in the kth cluster of the cell and the interfering channel from the lth adjacent cell BS to the cell-edge user as $\mathbf{H}_{k,\text{e}}$ and $\mathbf{G}_{k,l}$, respectively, we can express the received signal $\mathbf{y}_{k,\text{e}}$ at the cell-edge user as

$$\mathbf{y}_{k,\text{e}} = \mathbf{H}_{k,\text{e}}\mathbf{s}_k + \underbrace{\mathbf{H}_{k,\text{e}}\sum_{i=1,i\neq k}^{K}\mathbf{s}_i}_{\text{ICI}} + \underbrace{\sum_{l=1}^{L}\mathbf{G}_{k,l}\mathbf{x}_l}_{\text{MCI}} + \mathbf{n}_{k,\text{e}}, \tag{1}$$

where $\mathbf{x}_l$ represents the data vector transmitted by the lth adjacent cell BS and $\mathbf{n}_{k,\text{e}}$ indicates an AWGN vector with a mean zero vector and a covariance matrix $\sigma_n^2\mathbf{I}_N$. The data vector $\mathbf{s}_k$ is given by superimposing data streams for the users in the kth cluster according to the NOMA principle in the power domain and then transforming them by the transmit beamforming matrix $\mathbf{P}_k$ with size $M \times J$, which is expressed as

$$\begin{aligned} \mathbf{s}_k &= \mathbf{P}_k\mathbf{d}_k \\ \mathbf{d}_k &= \mathbf{D}_{k,\text{c}}\mathbf{d}_{k,\text{c}} + \mathbf{D}_{k,\text{e}}\mathbf{d}_{k,\text{e}}. \end{aligned} \tag{2}$$

$\mathbf{d}_{k,\text{c}}$ and $\mathbf{d}_{k,\text{e}}$ are the data symbol vectors for the cell-center and cell-edge users in the kth cluster, respectively. The power of each element in the data symbol vectors is normalized to unity. The $J \times J$ diagonal matrices $\mathbf{D}_{k,\text{c}}$ and $\mathbf{D}_{k,\text{e}}$ are the power allocation matrices for the cell-center and cell-edge users in the kth cluster, respectively, and $\sum_{k=1}^{K}\text{tr}\left\{\mathbf{D}_{k,\text{c}}^2 + \mathbf{D}_{k,\text{e}}^2\right\} = P_T$, where P_T is the total transmit power.

The cell-edge user applies the receive beamforming matrix $\mathbf{W}_{k,e}$ with size $N \times J$ to the received signal $\mathbf{y}_{k,e}$ in (1) in order to recover J desired data streams, as given in

$$\mathbf{W}_{k,e}^H \mathbf{y}_{k,e} = \mathbf{W}_{k,e}^H \mathbf{H}_{k,e} \mathbf{s}_k + \mathbf{W}_{k,e}^H \mathbf{H}_{k,e} \sum_{i=1,i\neq k}^{K} \mathbf{s}_i \\ + \mathbf{W}_{k,e}^H \sum_{l=1}^{L} \mathbf{G}_{k,l} \mathbf{x}_l + \mathbf{W}_{k,e}^H \mathbf{n}_{k,e}. \quad (3)$$

Since the MCI is negligible at cell-center users, the MCI is treated as merged into the AWGN at the users. Denoting the channel from the serving cell BS to the cell-center user in the kth cluster of the cell as $\mathbf{H}_{k,c}$, we can express the received signal $\mathbf{y}_{k,c}$ at the cell-center user as

$$\mathbf{y}_{k,c} = \mathbf{H}_{k,c} \mathbf{s}_k + \underbrace{\mathbf{H}_{k,c} \sum_{i=1,i\neq k}^{K} \mathbf{s}_i}_{\text{ICI}} + \mathbf{n}_{k,c}, \quad (4)$$

where $\mathbf{n}_{k,c}$ is an AWGN vector having the same attributes as $\mathbf{n}_{k,e}$. The cell-center user also applies the receive beamforming matrix $\mathbf{W}_{k,c}$ with size $N \times J$ to the received signal $\mathbf{y}_{k,c}$ in (4) in order to recover J desired data streams, as given in

$$\mathbf{W}_{k,c}^H \mathbf{y}_{k,c} = \mathbf{W}_{k,c}^H \mathbf{H}_{k,c} \mathbf{s}_k + \mathbf{W}_{k,c}^H \mathbf{H}_{k,c} \sum_{i=1,i\neq k}^{K} \mathbf{s}_i + \mathbf{W}_{k,c}^H \mathbf{n}_{k,c}. \quad (5)$$

3 Proposed MC-MIMO-NOMA technique for multi-cell systems

Figure 2 shows the transmitter structure of the BS for the proposed multi-cell MC-MIMO-NOMA technique and Fig. 3 shows the receiver structure of the cell-center and cell-edge users for the proposed technique. The transmitter sequentially performs power allocation $(\mathbf{D}_{k,c}, \mathbf{D}_{k,e})$, NOMA modulation, pre-processing $(\mathbf{Z}_k^{-H})$, and transmit beamforming $(\mathbf{P}_k)$. The receiver of the cell-center user sequentially performs receive beamforming $(\mathbf{W}_{k,c})$, post-processing $(\mathbf{U}_{k,c})$, SIC, and symbol detection, whereas the receiver of the cell-edge user sequentially performs receive beamforming $(\mathbf{W}_{k,e})$, post-processing $(\mathbf{U}_{k,e})$, and symbol detection.

The transmit and receive beamforming in this system is designed to mitigate the MCI and to eliminate the ICI at the cell-center and cell-edge users in order to improve spectral efficiency. In addition, this system performs the inter-cluster and inter-stream power allocation along with the pre-processing at the BS and the post-processing at the users to enhance spectral efficiency further. In what

follows, we present how to design the transmit and receive beamforming and how to determine the power allocation including the pre-processing at the BS and the post-processing at the users.

3.1 Design of transmit and receive beamforming matrices

The proposed transmit and receive beamforming matrices are designed in three steps. The first step is to find the receive beamforming matrices $\mathbf{W}_{1,e}, \mathbf{W}_{2,e}, \ldots, \mathbf{W}_{K,e}$ for K cell-edge users to mitigate the MCI effectively. That is, $\mathbf{W}_{k,e}$ is designed to minimize the MCI power and whiten the MCI plus noise as follows. First, the $N \times J$ matrix $\bar{\mathbf{W}}_{k,e}$ minimizing the MCI power is obtained from

$$\bar{\mathbf{W}}_{k,e} = \arg \min_{\bar{\mathbf{W}}_{k,e}^H \bar{\mathbf{W}}_{k,e} = \mathbf{I}_J} \text{tr}\left\{ \bar{\mathbf{W}}_{k,e}^H \mathbf{R}_{k,\text{MCI}} \bar{\mathbf{W}}_{k,e} \right\}, \quad (6)$$

where $\mathbf{R}_{k,\text{MCI}}$ is the MCI covariance matrix given as $\mathbf{R}_{k,\text{MCI}} = \sum_{l=1}^{L} \mathbf{G}_{k,l} E\{\mathbf{x}_l \mathbf{x}_l^H\} \mathbf{G}_{k,l}^H$. Solving (6), we find $\bar{\mathbf{W}}_{k,e}$ whose columns are composed of the eigenvectors of $\mathbf{R}_{k,\text{MCI}}$ corresponding to the J smallest eigenvalues $\lambda_{k,1}, \lambda_{k,2}, \ldots, \lambda_{k,J}$. Then, the matrix $\mathbf{W}_{k,e}$ whitening the MCI plus noise is given as

$$\mathbf{W}_{k,e} = \bar{\mathbf{W}}_{k,e} \cdot \text{diag}\left\{ \frac{1}{\sqrt{\lambda_{k,1} + \sigma_n^2}}, \frac{1}{\sqrt{\lambda_{k,2} + \sigma_n^2}}, \ldots, \frac{1}{\sqrt{\lambda_{k,J} + \sigma_n^2}} \right\}. \quad (7)$$

We note that the cell-edge users can find their own receive beamforming matrices without channel feedback. This reduces system overhead.

The second step is to use $\mathbf{W}_{1,e}, \mathbf{W}_{2,e}, \ldots, \mathbf{W}_{K,e}$ to find the transmit beamforming matrices $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_K$ that eliminate the ICI between the cell-edge users. In detail, by utilizing the effective channels $\mathbf{W}_{1,e}^H \mathbf{H}_{1,e}, \mathbf{W}_{2,e}^H \mathbf{H}_{2,e}, \ldots, \mathbf{W}_{K,e}^H \mathbf{H}_{K,e}$ to K cell-edge users in all the clusters, the $(K-1)J \times M$ matrices $\mathcal{H}_k$ for $1 \leq k \leq K$ are formed as

$$\mathcal{H}_{k,e} = \big[\mathbf{H}_{1,e}^H \mathbf{W}_{1,e} \ \mathbf{H}_{2,e}^H \mathbf{W}_{2,e} \ \cdots \ \mathbf{H}_{k-1,e}^H \mathbf{W}_{k-1,e} \\ \mathbf{H}_{k+1,e}^H \mathbf{W}_{k+1,e} \ \mathbf{H}_{k+2,e}^H \mathbf{W}_{k+2,e} \ \cdots \ \mathbf{H}_{K,e}^H \mathbf{W}_{K,e} \big]^H. \quad (8)$$

To eliminate the ICI, the following condition is required for $1 \leq k \leq K$.

$$\mathcal{H}_{k,e} \mathbf{P}_k = \mathbf{0}. \quad (9)$$

If the condition $M \geq KJ$ is satisfied, a nonzero matrix $\mathbf{P}_k$ satisfying (9) is given as

$$\mathbf{P}_k = \bar{\mathbf{V}}_k \boldsymbol{\Omega}_k. \quad (10)$$

Here, $\bar{\mathbf{V}}_k$ is the $M \times (M - KJ + J)$ matrix whose columns are composed of $(M - KJ + J)$ right singular vectors of

Fig. 2 Transmitter structure for a serving BS in the proposed multi-cell MC-MIMO-NOMA technique
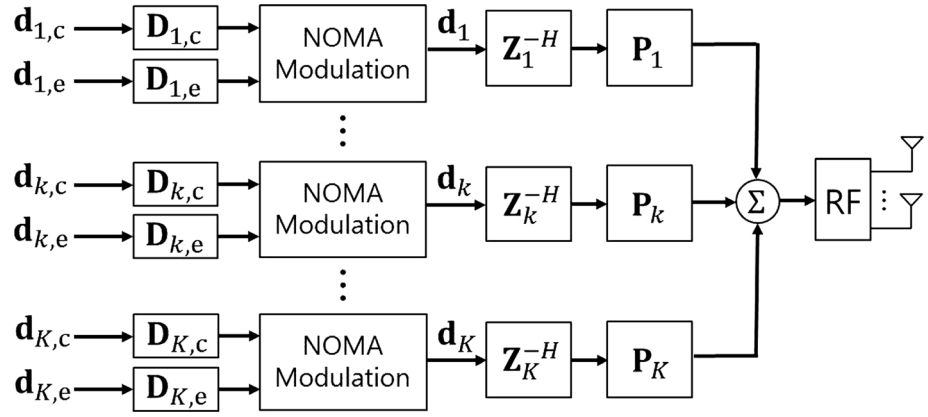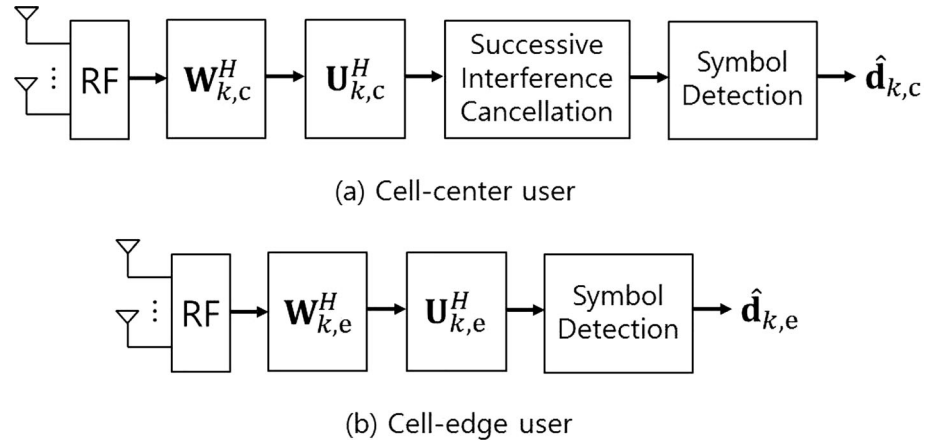

Fig. 3 Receiver structure for cell-center and cell-edge users in the proposed multi-cell MC-MIMO-NOMA technique

(a) Cell-center user

(b) Cell-edge user

$\mathcal{H}_{k,\mathrm{e}}$ corresponding to zero singular values and the $(M - KJ + J) \times J$ matrix $\boldsymbol{\Omega}_k$ is an arbitrary full rank matrix.

The matrix $\boldsymbol{\Omega}_k$ can be optimized to maximize the lower bound for the received SNR at the cell-edge user with the constraint $\|\boldsymbol{\Omega}_k(:,j)\| = 1$ for $1 \leq j \leq J$. Representing $\mathbf{W}_{k,\mathrm{e}}^H \mathbf{y}_{k,\mathrm{e}}$ in (3) using $\mathbf{W}_{k,\mathrm{e}}$ in (7) and $\mathbf{P}_k$ in (10) as

$$\mathbf{W}_{k,\mathrm{e}}^H \mathbf{y}_{k,\mathrm{e}} = \mathbf{W}_{k,\mathrm{e}}^H \mathbf{H}_{k,\mathrm{e}} \bar{\mathbf{V}}_k \boldsymbol{\Omega}_k \mathbf{d}_k + \mathbf{W}_{k,\mathrm{e}}^H \left(\sum_{l=1}^{L} \mathbf{G}_{k,l} \mathbf{x}_l + \mathbf{n}_{k,\mathrm{e}} \right), \tag{11}$$

we can write the received SNR at the cell-edge user, which is denoted by $\mathrm{SNR}_{k,\mathrm{e}}$, as

$$\mathrm{SNR}_{k,\mathrm{e}} = J^{-1} \, \mathrm{tr} \left\{ \mathbf{W}_{k,\mathrm{e}}^H \mathbf{H}_{k,\mathrm{e}} \bar{\mathbf{V}}_k \boldsymbol{\Omega}_k \mathbf{R}_{\mathbf{d}_k} \boldsymbol{\Omega}_k^H \bar{\mathbf{V}}_k^H \mathbf{H}_{k,\mathrm{e}}^H \mathbf{W}_{k,\mathrm{e}} \right\}, \tag{12}$$

where $\mathbf{R}_{\mathbf{d}_k} = E\{\mathbf{d}_k \mathbf{d}_k^H\}$. When we express $\mathbf{W}_{k,\mathrm{e}}^H \mathbf{H}_{k,\mathrm{e}} \bar{\mathbf{V}}_k$ by SVD with a $J \times J$ unitary matrix $\mathcal{U}_{k,\mathrm{e}}$, a $J \times J$ diagonal matrix $\boldsymbol{\Lambda}_{k,\mathrm{e}}$ and a $(M - KJ + J) \times J$ matrix $\mathcal{V}_{k,\mathrm{e}}$ as

$$\mathbf{W}_{k,\mathrm{e}}^H \mathbf{H}_{k,\mathrm{e}} \bar{\mathbf{V}}_k = \mathcal{U}_{k,\mathrm{e}} \boldsymbol{\Lambda}_{k,\mathrm{e}} \mathcal{V}_{k,\mathrm{e}}^H, \tag{13}$$

$\mathrm{SNR}_{k,\mathrm{e}}$ is lower-bounded as

$$\begin{aligned} \mathrm{SNR}_{k,\mathrm{e}} &= J^{-1} \, \mathrm{tr} \left\{ \boldsymbol{\Lambda}_{k,\mathrm{e}}^2 \mathcal{V}_{k,\mathrm{e}}^H \boldsymbol{\Omega}_k \mathbf{R}_{\mathbf{d}_k} \boldsymbol{\Omega}_k^H \mathcal{V}_{k,\mathrm{e}} \right\} \\ &\geq \det \left(\boldsymbol{\Lambda}_{k,\mathrm{e}}^2 \mathcal{V}_{k,\mathrm{e}}^H \boldsymbol{\Omega}_k \mathbf{R}_{\mathbf{d}_k} \boldsymbol{\Omega}_k^H \mathcal{V}_{k,\mathrm{e}} \right)^{\frac{1}{J}} \\ &= \left(\left| \det \left(\mathcal{V}_{k,\mathrm{e}}^H \boldsymbol{\Omega}_k \right) \right|^2 \det \left(\boldsymbol{\Lambda}_{k,\mathrm{e}}^2 \mathbf{R}_{\mathbf{d}_k} \right) \right)^{\frac{1}{J}}. \end{aligned} \tag{14}$$

From Hadamard's inequality that states that $|\det(\mathbf{A})| \leq \prod_{j=1}^{J} \|\mathbf{A}(:,j)\|$, we obtain

$$\left| \det \left(\mathcal{V}_{k,\mathrm{e}}^H \boldsymbol{\Omega}_k \right) \right| \leq \prod_{j=1}^{J} \left\| \mathcal{V}_{k,\mathrm{e}}^H \boldsymbol{\Omega}_k(:,j) \right\|. \tag{15}$$

It is found that to achieve the equality in (15) while maximizing $\prod_{j=1}^{J} \left\| \mathcal{V}_{k,\mathrm{e}}^H \boldsymbol{\Omega}_k(:,j) \right\|$, $\boldsymbol{\Omega}_k$ should be given as

$$\boldsymbol{\Omega}_k = \mathcal{V}_{k,\mathrm{e}} \mathbf{V}_{\Omega}, \tag{16}$$

where $\mathbf{V}_{\Omega}$ is an arbitrary unitary matrix with size $J \times J$. Thus, the transmit beamforming matrix $\mathbf{P}_k$ is obtained with $\mathbf{V}_{\Omega} = \mathbf{I}_J$ as

$$\mathbf{P}_k = \bar{\mathbf{V}}_k \mathcal{V}_{k,\mathrm{e}}. \tag{17}$$

The third step is to find the receive beamforming matrices

$\mathbf{W}_{1,c}, \mathbf{W}_{2,c}, \ldots, \mathbf{W}_{K,c}$ for K cell-center users to eliminate the ICI between the cell-center users. When we define the $N \times (K-1)J$ matrix $\mathcal{H}_{k,c}$ as

$$\mathcal{H}_{k,c} := \begin{bmatrix} \mathbf{H}_{k,c}\mathbf{P}_1 & \mathbf{H}_{k,c}\mathbf{P}_2 & \cdots & \mathbf{H}_{k,c}\mathbf{P}_{(k-1)} & \mathbf{H}_{k,c}\mathbf{P}_{(k+1)} \\ & \mathbf{H}_{k,c}\mathbf{P}_{(k+2)} & \cdots & \mathbf{H}_{k,c}\mathbf{P}_K \end{bmatrix}, \quad (18)$$

the ICI elimination requires

$$\mathbf{W}_{k,c}^H \mathcal{H}_{k,c} = \mathbf{0}, \quad k = 1, 2, \ldots, K. \quad (19)$$

If the condition $N \geq KJ$ is satisfied, a nonzero matrix $\mathbf{W}_{k,c}$ satisfying (19) is given as

$$\mathbf{W}_{k,c} = \bar{\mathbf{U}}_{k,c} \mathbf{\Psi}_k. \quad (20)$$

Here, $\bar{\mathbf{U}}_{k,c}$ is the $N \times (N - KJ + J)$ matrix whose columns are composed of $(N - KJ + J)$ left singular vectors of the matrix $\mathcal{H}_{k,c}$ corresponding to zero singular values and the $(N - KJ + J) \times J$ matrix $\mathbf{\Psi}_k$ is an arbitrary full rank matrix.

The matrix $\mathbf{\Psi}_{k,c}$ can be also optimized to maximize the lower bound for the received SNR at the cell-center user with the constraint $\|\mathbf{\Psi}_k(:,j)\| = 1$ for $1 \leq j \leq J$. Representing $\mathbf{W}_{k,c}^H \mathbf{y}_{k,c}$ in (5) using $\mathbf{W}_{k,c}$ in (20) as

$$\mathbf{W}_{k,c}^H \mathbf{y}_{k,c} = \mathbf{\Psi}_k^H \bar{\mathbf{U}}_{k,c}^H \mathbf{H}_{k,c} \mathbf{P}_k \mathbf{d}_k + \mathbf{\Psi}_k^H \bar{\mathbf{U}}_{k,c}^H \mathbf{n}_{k,c}, \quad (21)$$

we can write the received SNR at the cell-center user, which is denoted by $\text{SNR}_{k,c}$, as

$$\text{SNR}_{k,c} = J^{-1} \sigma_n^{-2} \text{tr}\left\{ \mathbf{\Psi}_k^H \bar{\mathbf{U}}_{k,c}^H \mathbf{H}_{k,c} \mathbf{P}_k \mathbf{R}_{\mathbf{d}_k} \mathbf{P}_k^H \mathbf{H}_{k,c}^H \bar{\mathbf{U}}_{k,c} \mathbf{\Psi}_k \right\}. \quad (22)$$

When we express $\bar{\mathbf{U}}_{k,c}^H \mathbf{H}_{k,c} \mathbf{P}_k$ by SVD with a $(N - KJ + J) \times J$ matrix $\mathcal{U}_{k,c}$, a $J \times J$ diagonal matrix $\mathbf{\Lambda}_{k,c}$ and a $J \times J$ unitary matrix $\mathcal{V}_{k,c}$ as

$$\bar{\mathbf{U}}_{k,c}^H \mathbf{H}_{k,c} \mathbf{P}_k = \mathcal{U}_{k,c} \mathbf{\Lambda}_{k,c} \mathcal{V}_{k,c}^H, \quad (23)$$

$\text{SNR}_{k,c}$ is lower-bounded as

$$\begin{aligned} \text{SNR}_{k,c} &= J^{-1} \sigma_n^{-2} \text{tr}\left\{ \mathbf{\Psi}_k^H \mathcal{U}_{k,c} \mathbf{\Lambda}_{k,c} \mathcal{V}_{k,c}^H \mathbf{R}_{\mathbf{d}_k} \mathcal{V}_{k,c} \mathbf{\Lambda}_{k,c} \mathcal{U}_{k,c}^H \mathbf{\Psi}_k \right\} \\ &\geq \sigma_n^{-2} \det\left(\mathbf{\Psi}_k^H \mathcal{U}_{k,c} \mathbf{\Lambda}_{k,c} \mathcal{V}_{k,c}^H \mathbf{R}_{\mathbf{d}_k} \mathcal{V}_{k,c} \mathbf{\Lambda}_{k,c} \mathcal{U}_{k,c}^H \mathbf{\Psi}_k \right)^{\frac{1}{J}} \\ &= \sigma_n^{-2} \left(\left| \det\left(\mathcal{U}_{k,c}^H \mathbf{\Psi}_k \right) \right|^2 \det\left(\mathbf{\Lambda}_{k,c}^2 \mathbf{R}_{\mathbf{d}_k} \right) \right)^{\frac{1}{J}}. \end{aligned} \quad (24)$$

From Hadamard's inequality, we obtain

$$\left| \det\left(\mathcal{U}_{k,c}^H \mathbf{\Psi}_k \right) \right| \leq \prod_{j=1}^{J} \left\| \mathcal{U}_{k,c}^H \mathbf{\Psi}_k(:,j) \right\|. \quad (25)$$

Like in (15), it is found that to achieve the equality in (25) while maximizing $\prod_{j=1}^{J} \left\| \mathcal{U}_{k,c}^H \mathbf{\Psi}_k(:,j) \right\|$, $\mathbf{\Psi}_k$ should be given as

$$\mathbf{\Psi}_k = \mathcal{U}_{k,c} \mathbf{U}_\Psi, \quad (26)$$

where $\mathbf{U}_\Psi$ is an arbitrary unitary matrix with size $J \times J$. Thus, the receive beamforming matrix $\mathbf{W}_{k,c}$ is obtained with $\mathbf{U}_\Psi = \mathbf{I}_J$ as

$$\mathbf{W}_{k,c} = \bar{\mathbf{U}}_{k,c} \mathcal{U}_{k,c}. \quad (27)$$

The cell-center users can also find their own receive beamforming matrices without channel feedback, thereby reducing system overhead.

In summary, the proposed technique finds the receive beamforming matrices for the cell-center and cell-edge users and the transmit beamforming matrices for the serving BS as follows.

Step 1: The cell-edge user in the kth cluster obtains the MCI covariance matrix $\mathbf{R}_{k,\text{MCI}}$ and finds its own receive beamforming matrix $\mathbf{W}_{k,e}$ in (7) using $\bar{\mathbf{W}}_{k,e}$ whose columns are composed of the eigenvectors of $\mathbf{R}_{k,\text{MCI}}$ corresponding to the J smallest eigenvalues.

Step 2: The serving BS obtains the effective CSI $\mathbf{W}_{k,e}^H \mathbf{H}_{k,e}$ for $1 \leq k \leq K$ with channel feedback from the K cell-edge users and finds the transmit beamforming matrices $\mathbf{P}_k$ for $1 \leq k \leq K$ as given in (17).

Step 3: The cell-center user in the kth cluster obtains the precoded CSI $\mathbf{H}_{k,c} \mathbf{P}_i$ for $1 \leq i \leq K$ through channel estimation and finds its own receive beamforming matrix $\mathbf{W}_{k,c}$ as given in (27).

Accurate CSI feedback from the cell-edge users is required to design effective transmit beamforming matrices. In consideration of the system overhead caused by the channel feedback, this technique may be beneficially employed in slow fading channel environments where frequent channel feedback is unnecessary. When imperfect CSI due to delayed feedback of quantized CSI is used to design the transmit beamforming matrices, the ICI between the cell-edge users cannot be eliminated perfectly, thereby causing performance degradation. In the future, this technique will be extended to be robust against imperfect CSI.

3.2 Determination of power allocation

By using the transmit and receive beamforming matrices designed in Sect. 3.1, we can rewrite the received signals at the cell-center and cell-edge users in the k cluster as

$$\mathbf{W}_{k,\mathrm{c}}^{H}\mathbf{y}_{k,\mathrm{c}} = \mathbf{W}_{k,\mathrm{c}}^{H}\mathbf{H}_{k,\mathrm{c}}\mathbf{P}_{k}\mathbf{d}_{k} + \mathbf{W}_{k,\mathrm{c}}^{H}\mathbf{n}_{k,\mathrm{c}} \tag{28}$$

$$\mathbf{W}_{k,\mathrm{e}}^{H}\mathbf{y}_{k,\mathrm{e}} = \mathbf{W}_{k,\mathrm{e}}^{H}\mathbf{H}_{k,\mathrm{e}}\mathbf{P}_{k}\mathbf{d}_{k} + \mathbf{W}_{k,\mathrm{e}}^{H}\boldsymbol{\epsilon}_{k,\mathrm{e}}, \tag{29}$$

where $\boldsymbol{\epsilon}_{k,\mathrm{e}} = \sum_{l=1}^{L}\mathbf{G}_{k,l}\mathbf{x}_{l} + \mathbf{n}_{k,\mathrm{e}}$. Then, we can express $\mathbf{W}_{k,\mathrm{c}}^{H}\mathbf{H}_{k,\mathrm{c}}\mathbf{P}_{k}$ and $\mathbf{W}_{k,\mathrm{e}}^{H}\mathbf{H}_{k,\mathrm{e}}\mathbf{P}_{k}$ by GSVD as, respectively,

$$\mathbf{W}_{k,\mathrm{c}}^{H}\mathbf{H}_{k,\mathrm{c}}\mathbf{P}_{k} = \mathbf{U}_{k,\mathrm{c}}\boldsymbol{\Sigma}_{k,\mathrm{c}}\bar{\mathbf{Z}}_{k}^{H} \tag{30}$$

$$\mathbf{W}_{k,\mathrm{e}}^{H}\mathbf{H}_{k,\mathrm{e}}\mathbf{P}_{k} = \mathbf{U}_{k,\mathrm{e}}\boldsymbol{\Sigma}_{k,\mathrm{e}}\bar{\mathbf{Z}}_{k}^{H}, \tag{31}$$

where $\mathbf{U}_{k,\mathrm{c}}$ and $\mathbf{U}_{k,\mathrm{e}}$ are $J \times J$ unitary matrices, $\boldsymbol{\Sigma}_{k,\mathrm{c}}$ and $\boldsymbol{\Sigma}_{k,\mathrm{e}}$ are $J \times J$ diagonal matrices, and $\bar{\mathbf{Z}}_{k}$ is a $J \times J$ non-singular matrix.

When we define the column-normalizing matrix $\boldsymbol{\Xi}_{k}$ as

$$\boldsymbol{\Xi}_{k} := \mathrm{diag}\Big\{\big\|\bar{\mathbf{Z}}_{k}^{-H}(:,1)\big\|^{-1}, \big\|\bar{\mathbf{Z}}_{k}^{-H}(:,2)\big\|^{-1}, \ldots, \big\|\bar{\mathbf{Z}}_{k}^{-H}(:,J)\big\|^{-1}\Big\}, \tag{32}$$

the BS applies the pre-processing matrix $\mathbf{Z}_{k}^{-H} = \bar{\mathbf{Z}}_{k}^{-H}\boldsymbol{\Xi}_{k}$ to the data symbol vector $\mathbf{d}_{k}$ prior to the transmit beamforming $\mathbf{P}_{k}$ as shown in Fig. 2. The cell-center and cell-edge users apply the post-processing matrices $\mathbf{U}_{k,\mathrm{c}}^{H}$ and $\mathbf{U}_{k,\mathrm{e}}^{H}$ to their receive beamformed signals $\mathbf{W}_{k,\mathrm{c}}^{H}\mathbf{y}_{k,\mathrm{c}}$ and $\mathbf{W}_{k,\mathrm{e}}^{H}\mathbf{y}_{k,\mathrm{e}}$, respectively, which is shown in Fig. 3. Then, the received signals $\bar{\mathbf{y}}_{k,\mathrm{c}} = \mathbf{U}_{k,\mathrm{c}}^{H}\mathbf{W}_{k,\mathrm{c}}^{H}\mathbf{y}_{k,\mathrm{c}}$ and $\bar{\mathbf{y}}_{k,\mathrm{e}} = \mathbf{U}_{k,\mathrm{e}}^{H}\mathbf{W}_{k,\mathrm{e}}^{H}\mathbf{y}_{k,\mathrm{e}}$ are expressed as

$$\begin{aligned}\bar{\mathbf{y}}_{k,\mathrm{c}} = \mathbf{U}_{k,\mathrm{c}}^{H}\mathbf{W}_{k,\mathrm{c}}^{H}\mathbf{y}_{k,\mathrm{c}} &= \boldsymbol{\Sigma}_{k,\mathrm{c}}\boldsymbol{\Xi}_{k}\mathbf{d}_{k} + \mathbf{U}_{k,\mathrm{c}}^{H}\mathbf{W}_{k,\mathrm{c}}^{H}\mathbf{n}_{k,\mathrm{c}} \\ &= \bar{\boldsymbol{\Sigma}}_{k,\mathrm{c}}\mathbf{d}_{k} + \bar{\mathbf{n}}_{k,\mathrm{c}}\end{aligned} \tag{33}$$

$$\begin{aligned}\bar{\mathbf{y}}_{k,\mathrm{e}} = \mathbf{U}_{k,\mathrm{e}}^{H}\mathbf{W}_{k,\mathrm{e}}^{H}\mathbf{y}_{k,\mathrm{e}} &= \boldsymbol{\Sigma}_{k,\mathrm{e}}\boldsymbol{\Xi}_{k}\mathbf{d}_{k} + \mathbf{U}_{k,\mathrm{e}}^{H}\mathbf{W}_{k,\mathrm{e}}^{H}\boldsymbol{\epsilon}_{k,\mathrm{e}} \\ &= \bar{\boldsymbol{\Sigma}}_{k,\mathrm{e}}\mathbf{d}_{k} + \bar{\boldsymbol{\epsilon}}_{k,\mathrm{e}},\end{aligned} \tag{34}$$

where $\bar{\boldsymbol{\Sigma}}_{k,\mathrm{c}} = \boldsymbol{\Sigma}_{k,\mathrm{c}}\boldsymbol{\Xi}_{k}$, $\bar{\boldsymbol{\Sigma}}_{k,\mathrm{e}} = \boldsymbol{\Sigma}_{k,\mathrm{e}}\boldsymbol{\Xi}_{k}$, $\bar{\mathbf{n}}_{k,\mathrm{c}} = \mathbf{U}_{k,\mathrm{c}}^{H}\mathbf{W}_{k,\mathrm{c}}^{H}\mathbf{n}_{k,\mathrm{c}}$, and $\bar{\boldsymbol{\epsilon}}_{k,\mathrm{e}} = \mathbf{U}_{k,\mathrm{e}}^{H}\mathbf{W}_{k,\mathrm{e}}^{H}\boldsymbol{\epsilon}_{k,\mathrm{e}}$. Thus, the MIMO-NOMA system in each cluster is decomposed into J parallel SISO-NOMA systems, which enables the application of existing techniques for SISO-NOMA systems.

Next, we discuss power allocation maximizing the sum rate of this system. Denoting the rate at the cell-center user and the rate at the cell-edge user in the kth cluster as $\mathcal{R}_{k,\mathrm{c}}$ and $\mathcal{R}_{k,\mathrm{e}}$, respectively, we can express $\mathcal{R}_{k,\mathrm{c}}$ and $\mathcal{R}_{k,\mathrm{e}}$ achievable by the system with perfect SIC at the cell-center user as

$$\begin{aligned}\mathcal{R}_{k,\mathrm{c}} &= \sum_{j=1}^{J}\log_{2}\big(1 + h_{k,j}p_{k,j}\big), \\ \mathcal{R}_{k,\mathrm{e}} &= \sum_{j=1}^{J}\log_{2}\bigg(1 + \frac{f_{k,j}q_{k,j}}{f_{k,j}p_{k,j}+1}\bigg),\end{aligned} \tag{35}$$

where $h_{k,j} = \frac{\big|\bar{\Sigma}_{k,\mathrm{c}}(j,j)\big|^{2}}{\sigma_{n}^{2}}$, $f_{k,j} = \min\Big\{\big|\bar{\Sigma}_{k,\mathrm{e}}(j,j)\big|^{2}, \frac{\big|\bar{\Sigma}_{k,\mathrm{c}}(j,j)\big|^{2}}{\sigma_{n}^{2}}\Big\}$, $p_{k,j} = \mathbf{D}_{k,\mathrm{c}}^{2}(j,j)$ and $q_{k,j} = \mathbf{D}_{k,\mathrm{e}}^{2}(j,j)$. To maximize the sum rate $\mathcal{R}_{\mathrm{sum}}$, which is given as $\mathcal{R}_{\mathrm{sum}} = \sum_{k=1}^{K}\big(\mathcal{R}_{k,\mathrm{c}} + \mathcal{R}_{k,\mathrm{e}}\big)$, with a fair power allocation between clusters as well as between users in each cluster, the power allocation for $p_{k,j}$ and $q_{k,j}$ is optimized as follows:

$$\big\{\mathbf{p}^{\star},\ \mathbf{q}^{\star}\big\} = \arg\max_{\mathbf{p},\ \mathbf{q}} \mathcal{R}_{\mathrm{sum}}$$

subject to:

$$\begin{aligned}&\sum_{k=1}^{K}\bar{P}_{k} \le P_{T},\ \bar{P}_{k} \ge v_{k}\frac{P_{T}}{K}, \\ &\sum_{j=1}^{J}p_{k,j} \le \delta_{k}\bar{P}_{k}, \quad \text{and} \quad \sum_{j=1}^{J}q_{k,j} \le (1-\delta_{k})\bar{P}_{k},\end{aligned} \tag{36}$$

where $\mathbf{p} = [p_{1,1}, p_{1,2}, \ldots, p_{1,J}, \ldots, p_{K,1}, p_{K,2}, \ldots, p_{K,J}]$, $\mathbf{q} = [q_{1,1}, q_{1,2}, \ldots, q_{1,J}, \ldots, q_{K,1}, q_{K,2}, \ldots, q_{K,J}]$, and $\bar{P}_{k}$ is the power to be allocated to the kth cluster. The parameters v_{k} and δ_{k} are all set to values between 0 and 1 for fairness between users in each cluster as well as clusters. This optimization is performed by alternately employing two processes described below. In the first process, the optimal power allocation for clusters is initially estimated. In the second process, the optimal powers to be allocated to the cell-center and cell-edge users in each cluster are determined using the power allocated to the cluster in the first process. These processes are iteratively performed until $\bar{P}_{k}$, $p_{k,j}$ and $q_{k,j}$ are all obtained with the required accuracy.

In detail, to estimate the power allocation for clusters maximizing the sum rate in the first process, we express $p_{k,j}$ and $q_{k,j}$ to $a_{k,j}\bar{P}_{k}$ and $b_{k,j}\bar{P}_{k}$, respectively. $a_{k,j}$ and $b_{k,j}$ indicate the ratios of $\bar{P}_{k}$ allocated to the cell-center and cell-edge users in the kth cluster, respectively. In the first process, we find $\bar{P}_{k}^{(m)}$, which is the power allocated to the kth cluster at the mth iteration, by maximizing the function $\mathcal{L}^{(m)}$ as

$$\begin{aligned}\mathcal{L}^{(m)}\big(\bar{\mathbf{P}}^{(m)},\ \mu^{(m)}\big) &= \sum_{k=1}^{K}\sum_{j=1}^{J}\Bigg\{\log_{2}\Big(1 + h_{k,j}a_{k,j}^{(m)}\bar{P}_{k}^{(m)}\Big) \\ &+ \log_{2}\bigg(1 + \frac{f_{k,j}b_{k,j}^{(m)}\bar{P}_{k}^{(m)}}{f_{k,j}a_{k,j}^{(m)}\bar{P}_{k}^{(m)}+1}\bigg)\Bigg\} + \mu^{(m)}\bigg(P_{T} - \sum_{k=1}^{K}\bar{P}_{k}^{(m)}\bigg)\end{aligned} \tag{37}$$

$$\frac{\partial \mathcal{L}^{(m)}}{\partial \bar{P}_k^{(m)}} = \sum_{j=1}^{J} \left\{ \frac{\alpha_{k,j}^{(m)}}{\alpha_{k,j}^{(m)} \bar{P}_k^{(m)} + 1} + \frac{\beta_{k,j}^{(m)}}{\beta_{k,j}^{(m)} \bar{P}_k^{(m)} + 1} - \frac{\gamma_{k,j}^{(m)}}{\gamma_{k,j}^{(m)} \bar{P}_k^{(m)} + 1} \right\}$$
$$- \mu^{(m)} = 0$$
(38)

$$\bar{P}_k^{(m)} = \max \left\{ \tilde{P}_k^{(m)}, \ v_k \frac{P_T}{K} \right\},$$
(39)

where $\bar{\mathbf{P}}^{(m)} = \left[\bar{P}_1^{(m)}, \bar{P}_2^{(m)}, \dots, \bar{P}_K^{(m)} \right]$, $\alpha_{k,j}^{(m)} = h_{k,j} a_{k,j}^{(m)}$, $\beta_{k,j}^{(m)} = f_{k,j} \left(a_{k,j}^{(m)} + b_{k,j}^{(m)} \right)$ and $\gamma_{k,j}^{(m)} = f_{k,j} a_{k,j}^{(m)}$. $\mu^{(m)}$ is a Lagrange multiplier in the mth iteration. $a_{k,j}^{(m)}$ and $b_{k,j}^{(m)}$ indicate the ratios of $\bar{P}_k^{(m)}$ allocated to the cell-center and cell-edge users in the kth cluster at the mth iteration, respectively. Furthermore, $a_{k,j}^{(m)}$ and $b_{k,j}^{(m)}$ are refined in the second process and transferred to the first process for the next iteration, which will be described later. $\tilde{P}_k^{(m)}$ in (39) indicates the value of $\bar{P}_k^{(m)}$ obtained from (38) for a given value of $\mu^{(m)}$. Equations (38) and (39) are repeatedly solved for $\bar{P}_k^{(m)}$ by adjusting the value of $\mu^{(m)}$ until the sum power $\sum_{k=1}^{K} \bar{P}_k^{(m)}$ is sufficiently close to P_T.

In the second process, we find $p_{k,j}^{(m)}$ and $q_{k,j}^{(m)}$, which are the estimates of $p_{k,j}$ and $q_{k,j}$ at the mth iteration, maximizing the function $l_k^{(m)}$ with $\bar{P}_k^{(m)}$ obtained from the first process as follows:

$$l_k^{(m)} \left(\mathbf{p}_k^{(m)}, \ \mathbf{q}_k^{(m)}, \ \mu_{k,c}^{(m)}, \ \mu_{k,e}^{(m)} \right)$$
$$= \sum_{j=1}^{J} \left\{ \log_2 \left(1 + h_{k,j} p_{k,j}^{(m)} \right) + \log_2 \left(1 + \frac{f_{k,j} q_{k,j}^{(m)}}{f_{k,j} p_{k,j}^{(m)} + 1} \right) \right\}$$
$$+ \mu_{k,c}^{(m)} \left(\delta_k \bar{P}_k^{(m)} - \sum_{j=1}^{J} p_{k,j}^{(m)} \right)$$
$$+ \mu_{k,e}^{(m)} \left((1 - \delta_k) \bar{P}_k^{(m)} - \sum_{j=1}^{J} q_{k,j}^{(m)} \right)$$
(40)

$$\frac{\partial l_k^{(m)}}{\partial q_{k,j}^{(m)}} = \frac{f_{k,j}}{f_{k,j} q_{k,j}^{(m)} + f_{k,j} p_{k,j}^{(m)} + 1} - \mu_{k,e}^{(m)} = 0$$
(41)

$$\frac{\partial l_k^{(m)}}{\partial p_{k,j}^{(m)}} = \frac{h_{k,j}}{h_{k,j} p_{k,j}^{(m)} + 1} + \frac{f_{k,j}}{f_{k,j} p_{k,j}^{(m)} + f_{k,j} q_{k,j}^{(m)} + 1}$$
$$- \frac{f_{k,j}}{f_{k,j} p_{k,j}^{(m)} + 1} - \mu_{k,c}^{(m)} = 0,$$
(42)

where $\mathbf{p}_k^{(m)} = \left[p_{k,1}^{(m)}, p_{k,2}^{(m)}, \dots, p_{k,J}^{(m)} \right]$ and $\mathbf{q}_k^{(m)} = \left[q_{k,1}^{(m)}, q_{k,2}^{(m)}, \dots, q_{k,J}^{(m)} \right]$. $\mu_{k,c}^{(m)}$ and $\mu_{k,e}^{(m)}$ are Lagrange

multipliers in the mth iteration. The optimal $p_{k,j}^{(m)}$ and $q_{k,j}^{(m)}$ in (41) and (42) are obtained using the alternating maximization in [29]. When we denote the values of $p_{k,j}^{(m)}$ and $q_{k,j}^{(m)}$ at the ith alternation of the mth iteration as $p_{k,j}^{(m(i))}$ and $q_{k,j}^{(m(i))}$, respectively, the maximization is performed as

$$q_{k,j}^{(m(i))} = \max \left\{ \frac{1}{\mu_{k,e}^{(m)}} - p_{k,j}^{(m(i-1))} - \frac{1}{f_{k,j}}, \ 0 \right\}$$
(43)

$$\frac{h_{k,j}}{h_{k,j} p_{k,j}^{(m(i))} + 1} + \frac{f_{k,j}}{f_{k,j} p_{k,j}^{(m(i))} + f_{k,j} q_{k,j}^{(m(i))} + 1}$$
$$- \frac{f_{k,j}}{f_{k,j} p_{k,j}^{(m(i))} + 1} - \mu_{k,c}^{(m)} = 0$$
(44)

$$p_{k,j}^{(m(i))} = \max \left\{ \tilde{p}_{k,j}^{(m(i))}, \ 0 \right\},$$
(45)

where $\tilde{p}_{k,j}^{(m(i))}$ in (45) represents the value of $p_{k,j}^{(m(i))}$ obtained from (44) for a given value of $\mu_{k,c}^{(m)}$. Since $p_{k,j}^{(m(i))}$ is not available until the ith alternation is completed, $p_{k,j}^{(m(i-1))}$ instead of $p_{k,j}^{(m(i))}$ is used to find $q_{k,j}^{(m(i))}$ in (43). The alternating maximization in (43)–(45) is continued until the sum powers $\sum_{j=1}^{J} p_{k,j}^{(m(i))}$ and $\sum_{j=1}^{J} q_{k,j}^{(m(i))}$ are sufficiently close to $\delta_k \bar{P}_k^{(m)}$ and $(1 - \delta_k) \bar{P}_k^{(m)}$ by adjusting the values of $\mu_{k,c}^{(m)}$ and $\mu_{k,e}^{(m)}$, respectively.

When the alternating maximization in the mth iteration is completed at the Ith alternation, $p_{k,j}^{(m(I))}$ and $q_{k,j}^{(m(I))}$ are assigned to $p_{k,j}^{(m)}$ and $q_{k,j}^{(m)}$, respectively. Then, the power ratios $a_{k,j}^{(m+1)}$ and $b_{k,j}^{(m+1)}$ for the $(m + 1)$th iteration in the first process are refined as

$$a_{k,j}^{(m+1)} = \frac{p_{k,j}^{(m)}}{\sum_{j=1}^{J} \left(p_{k,j}^{(m)} + q_{k,j}^{(m)} \right)},$$
$$b_{k,j}^{(m+1)} = \frac{q_{k,j}^{(m)}}{\sum_{j=1}^{J} \left(p_{k,j}^{(m)} + q_{k,j}^{(m)} \right)}.$$
(46)

$a_{k,j}^{(m+1)}$ and $b_{k,j}^{(m+1)}$ are returned to the first process in order to find $\bar{P}_k^{(m+1)}$, which is in turn used to find $p_{k,j}^{(m+1)}$ and $q_{k,j}^{(m+1)}$ in the second process. These processes are alternately performed until $\bar{P}_k$, $p_{k,j}$ and $q_{k,j}$ are all obtained with the required accuracy. Finally, the power allocation matrices $\mathbf{D}_{k,c}$ and $\mathbf{D}_{k,e}$ for the kth cluster are given using $p_{k,j}$ and $q_{k,j}$ obtained from this optimization as

$$\mathbf{D}_{k,c} = \text{diag} \left\{ \sqrt{p_{k,1}}, \ \sqrt{p_{k,2}}, \dots, \sqrt{p_{k,J}} \right\}$$
$$\mathbf{D}_{k,e} = \text{diag} \left\{ \sqrt{q_{k,1}}, \ \sqrt{q_{k,2}}, \dots, \sqrt{q_{k,J}} \right\}.$$
(47)

In summary, the proposed procedure optimizing the power allocation is as follows:

Step 1: The index m is initialized to 1 and the power ratios $a_{k,j}^{(m)}$ and $b_{k,j}^{(m)}$ are set to initial values.

Step 2: The power $\bar{P}_k^{(m)}$ allocated to the kth cluster at the mth iteration is obtained using (38) and (39).

Step 3: The powers $p_{k,j}^{(m)}$ and $q_{k,j}^{(m)}$, which are allocated to the cell-center and cell-edge users in the kth cluster at the mth iteration, respectively, are obtained using (43)–(45) with the power constraint $\bar{P}_k^{(m)}$.

Step 4: If $\bar{P}_k$, $p_{k,j}$ and $q_{k,j}$ are all obtained with the required accuracy, this procedure is terminated and the power allocation matrices $\mathbf{D}_{k,\text{c}}$ and $\mathbf{D}_{k,\text{e}}$ are obtained as given in (47). Otherwise, Step 5 is performed.

Step 5: The power ratios $a_{k,j}^{(m+1)}$ and $b_{k,j}^{(m+1)}$ are refined using (46) and the index m is increased by 1. Then, Steps 2, 3, and 4 are re-performed sequentially.

Steps 1 and 2 represent the first process, while Steps 3, 4, and 5 represent the second process.

This power allocation is executed at BSs and may require high computational complexity because it iteratively finds roots of nonlinear equations. Thus, it is suitable for use in BSs with high computational capabilities. In this power allocation, as the number of clusters (K) increases, the complexity of the first process allocating power to K clusters increases. And as the number of data streams transmitted to each cluster ($2J$) increases, the number of alternations in the second process tends to become larger to allocate power to the data streams with the required accuracy, thereby increasing computational complexity. When the value of KJ is fixed, the complexity of this power allocation becomes higher for a moderate number of clusters. In the future, it would be worth exploring non-iterative power allocation methods that achieve the performance close to the proposed power allocation to reduce computational complexity for use in low-cost BSs.

4 Simulation results

In this section, we evaluate the sum rate and outage performances of the proposed MC-MIMO-NOMA technique through simulations. In the simulations, the serving cell is considered to be interfered from three adjacent cells, which means $L = 3$. For the kth cluster, the distances between the serving BS and the cell-center user, between the serving BS
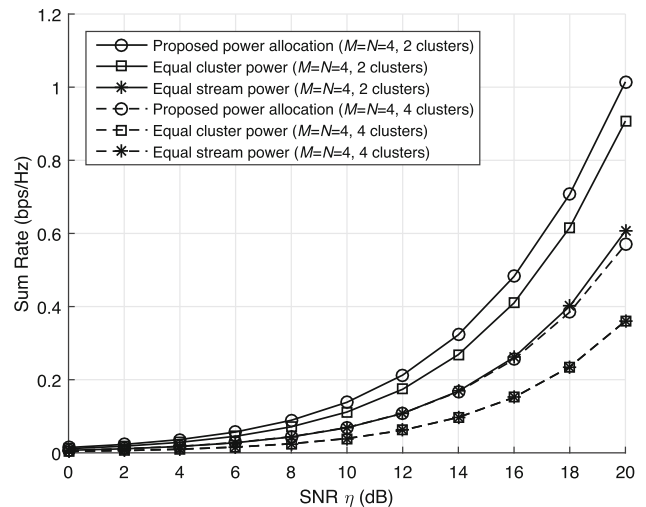

Fig. 4 Comparison of the sum rate performances according to different power allocation schemes applied to the proposed transmit and receive beamforming technique in the MC-MIMO-NOMA system with $M = N = 4$

and the cell-edge user, and between the lth adjacent cell BS and the cell-edge user are denoted as $r_{k,\text{c}}$, $r_{k,\text{e}}$, and $r_{k,l}$, respectively. We consider that while $r_{k,\text{c}}$ is uniformly distributed between 2m and 3m, $r_{k,\text{e}}$ and $r_{k,l}$ are uniformly distributed between 9m and 10m. BSs employ the number of transmit antennas (M) equal to the number of receive antennas (N) equipped at the users.

The channel matrices $\mathbf{H}_{k,\text{c}}$, $\mathbf{H}_{k,\text{e}}$, and $\mathbf{G}_{k,l}$ are given as $r_{k,\text{c}}^{-\frac{\alpha}{2}}\bar{\mathbf{H}}_{k,\text{c}}$, $r_{k,\text{e}}^{-\frac{\alpha}{2}}\bar{\mathbf{H}}_{k,\text{e}}$, and $r_{k,l}^{-\frac{\alpha}{2}}\bar{\mathbf{G}}_{k,l}$, respectively, where the elements of $\bar{\mathbf{H}}_{k,\text{c}}$, $\bar{\mathbf{H}}_{k,\text{e}}$, and $\bar{\mathbf{G}}_{k,l}$ are randomly generated from circular symmetric complex Gaussian distributions with zero mean and unit variance, and the path loss exponent α is set to 3. The serving BS transmits $\frac{M}{K}$ NOMA-modulated data streams to each cluster, where K indicates the number of clusters simultaneously supported by the serving BS. The parameter v_k is set to 0.2. This means that the serving BS allocates at least $0.2\frac{P_T}{K}$ of power to each cluster, where P_T indicates the total transmit power of the serving BS. The parameter δ_k is set to 0.1. This means that the serving BS may allocate up to $0.1\bar{P}_k$ of power to the cell-center user in the kth cluster, where $\bar{P}_k$ indicates the transmit power allocated to the kth cluster. The SNR η is given as $\eta = \frac{P_T}{\sigma_n^2}$. Each sum rate performance is found by averaging the instantaneous sum rates obtained from 2000 random channel realizations.

To show the impact of power allocation on the proposed transmit and receive beamforming technique for MC-MIMO-NOMA systems, Figs. 4 and 5 compare the sum rate performances obtained from three power allocation schemes. The schemes are the proposed power allocation, equal cluster power allocation, and equal stream power
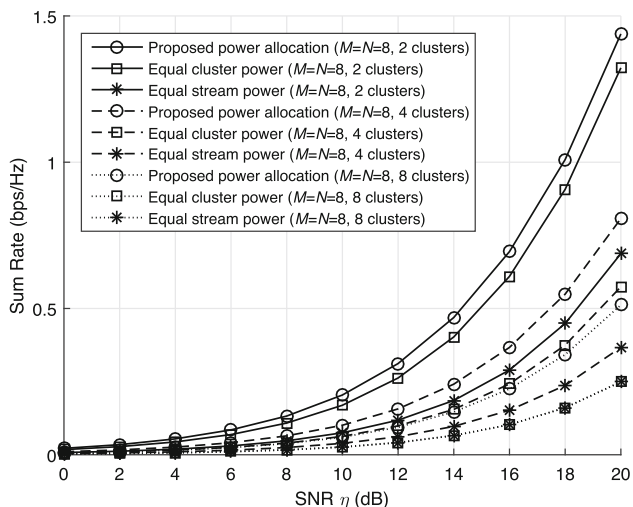
Fig. 5 Comparison of the sum rate performances according to different power allocation schemes applied to the proposed transmit and receive beamforming technique in the MC-MIMO-NOMA system with $M = N = 8$

allocation. Equal power allocation associated with the equal cluster power allocation and the equal stream power allocation is widely used as a baseline scheme for performance comparison [29, 33]. The equal cluster power allocation is the same as the proposed power allocation, except that the serving BS allocates $\frac{P_T}{K}$ of power to each cluster. The equal stream power allocation scheme evenly distributes the total transmit power P_T to M data streams. The MC-MIMO-NOMA systems with $M = N = 4$ and $M = N = 8$ are considered in Figs. 4 and 5, respectively.

As we can see from those figures, the propose power allocation scheme achieves the best performances of them. This is because the proposed scheme effectively allocates the transmit power for each cluster and the transmit power for each data stream in consideration of channel states in all clusters. In detail, as the number of clusters simultaneously supported by the serving BS decreases, the sum rate performances of the proposed scheme are improved further. When the number of clusters decreases, the transmit power to be allocated to each cluster increases on average and the number of data streams for each cluster increases as well. By effectively allocating more transmit power to data streams to be transmitted through good quality channels in each cluster, the proposed scheme improves the sum rate performances. In addition, as the number of clusters increases, the performance gaps between the proposed scheme and the others become larger. This is because the proposed scheme exploits the increased cluster diversity to allocate the transmit power to each cluster efficaciously. When the serving BS simultaneously supports as many clusters as the number of transmit antennas, only a single NOMA modulated data
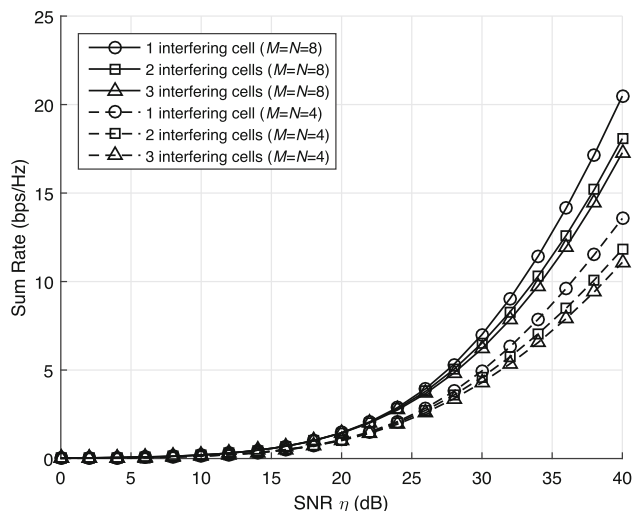

Fig. 6 Comparison of the sum rate performances of the proposed technique in the MC-MIMO-NOMA systems with $M = N = 4$ and $M = N = 8$ as the number of interfering cells increases from 1 to 3
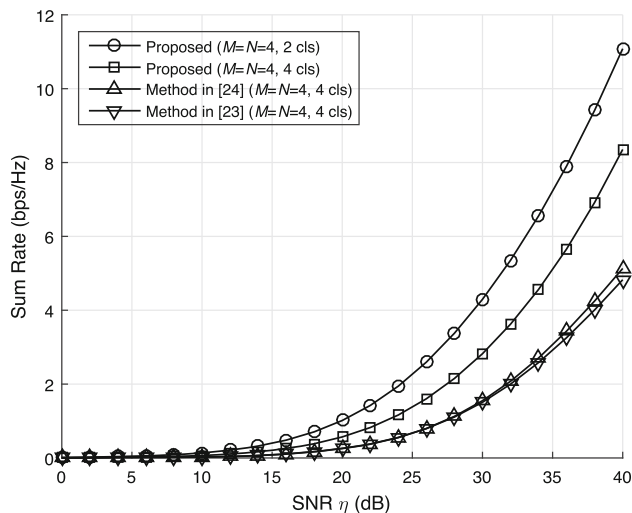

Fig. 7 Comparison of the sum rate performances between the proposed technique and the existing techniques in the MC-MIMO-NOMA system with $M = N = 4$

stream is transmitted to each cluster. Since the equal stream power allocation is identical to the equal cluster power allocation in this case, the performances of both schemes are identical in those figures.

Figure 6 demonstrates the sum rate performances of the proposed MC-MIMO-NOMA technique, which employs both the proposed transmit and receive beamforming and the proposed power allocation, with the increasing number of interfering cells. In this simulation, we consider two systems with $M = N = 4$ and $M = N = 8$ and the serving BS simultaneously supports 2 clusters in each system. It is shown in all the systems that regardless of the number of interfering cells, the proposed technique consistently
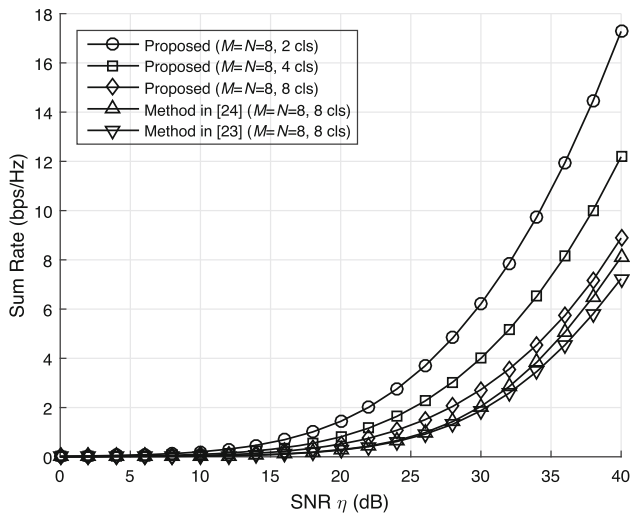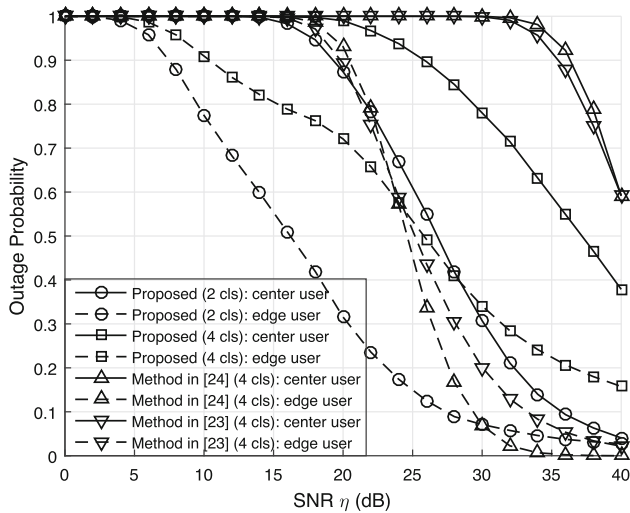
Fig. 8 Comparison of the sum rate performances between the proposed technique and the existing techniques in the MC-MIMO-NOMA system with $M = N = 8$
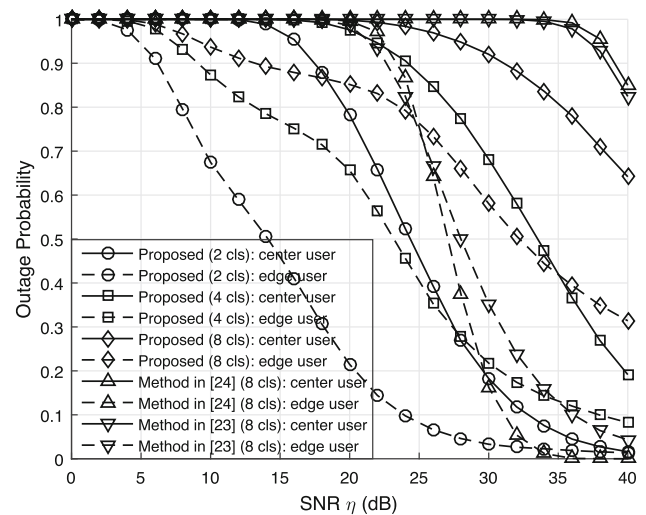

Fig. 10 Comparison of the outage performances for cell-center and cell-edge users between the proposed technique and the existing techniques in the MC-MIMO-NOMA system with $M = N = 8$

MIMO-NOMA systems with $M = N = 4$ and $M = N = 8$, respectively. We can see that regardless of the number of clusters simultaneously supported by the serving BS in all the systems, the proposed technique achieves better performances than the existing techniques. Since the existing techniques transmit only a single NOMA modulated data stream to each cluster, they may simultaneously support as many clusters as the number of transmit antennas. However, the existing techniques cannot be extend to transmit multiple data streams to each cluster even though they simultaneously support fewer clusters than the number of transmit antennas. In contrast, the proposed technique can transmit multiple data streams to each cluster at the expense of simultaneously supporting fewer clusters than the number of transmit antennas, thereby improving the sum rate performances. Thus, the proposed technique may provide a good balance between the number of clusters to be simultaneously supported by the serving BS and the sum rate performance by adjusting the number of data streams to be transmitted to each cluster as needed.

Figures 9 and 10 demonstrate the outage probability for a cell-center user, which is given as $\Pr(\mathcal{R}_{k,\mathrm{c}} < 1\,\mathrm{bps/Hz})$ and that for a cell-edge user, which is given as $\Pr(\mathcal{R}_{k,\mathrm{e}} < 0.02\,\mathrm{bps/Hz})$, in the kth cluster. The MC-MIMO-NOMA systems with $M = N = 4$ and $M = N = 8$ are considered in Figs. 9 and 10, respectively. It can be seen from both figures that regardless of the number of clusters, the proposed technique achieves lower outage probabilities for the cell-center user than the existing techniques. In addition, as the number of clusters simultaneously supported by the serving BS decreases, the outage performance for the cell-center user is further improved in the proposed technique. This is because the number of data

Fig. 9 Comparison of the outage performances for cell-center and cell-edge users between the proposed technique and the existing techniques in the MC-MIMO-NOMA system with $M = N = 4$

improves the performances as the SNR η increases. This confirms the effectiveness of the proposed technique. However, the performances gradually deteriorate with the increasing number of interfering cells. This is because the residual MCI power after MCI mitigation also increases as the number of interfering cells increases.

In Figs. 7 and 8, the sum rate performances of the proposed technique are compared with those of the existing MC-MIMO-NOMA techniques that are the MIMO-NOMA technique in [23] and the signal-alignment based MIMO-NOMA technique in [24]. The existing techniques utilize time sharing among multiple BSs to avoid MCI. Figures 7 and 8 show the performance comparisons for the MC-

streams transmitted to each cluster increases with the decreasing number of the clusters.

As we can see from Fig. 9, the outage probabilities for the cell-edge user in the proposed technique are lower than those in the existing techniques at SNRs less than 24 dB. Furthermore, the proposed technique with 2 clusters achieves better outage performance than the existing techniques at SNRs less than 30 dB. Figure 10 shows similar performance trends to Fig. 9. In detail, the outage probabilities for the cell-edge user in the proposed technique are lower than those in the existing techniques at SNRs less than 25 dB. While the proposed technique with 4 clusters achieves better outage performance than the existing techniques at SNRs less than 29 dB, the proposed technique with 2 clusters achieves better outage performance than the existing techniques at SNRs less than 33 dB. Since the number of data streams transmitted to each cluster increases with the decreasing number of the clusters in the proposed technique, the outage performance for the cell-edge user in the proposed technique is improved with the decreasing number of the clusters. In contrast, the existing techniques employing MCI avoidance achieve better outage performance for the cell-edge user than the proposed technique at SNRs greater than 30 or 33 dB as shown in Figs. 9 and 10. This is because the performance for the cell-edge user in the proposed technique is affected by the residual MCI power after MCI mitigation.

We see from this performance evaluation that the proposed technique can improve the outage performances for all the users by transmitting multiple data streams to each cluster at the expense of simultaneously supporting fewer clusters than the number of transmit antennas. Thus, the proposed technique may provide a good balance between the number of clusters to be simultaneously supported by the serving BS and the outage performances by adjusting the number of data streams to be transmitted to each cluster as needed. Considering comprehensively the sum rate, the outage probabilities for all the users, and the flexibility for balancing between the performances and the number of clusters simultaneously supported by the serving BS, we confirm that the proposed technique is more beneficial than the existing techniques.

To investigate the sum rate and outage performances of the proposed technique for SC-MIMO-NOMA systems, we compare the performances of the proposed technique transmitting data streams to only a single cluster with those of the existing SC-MIMO-NOMA techniques. The existing techniques are the SC-MIMO-NOMA technique exploiting GSVD and equal stream power allocation in [26, 27] and the SC-MIMO-NOMA technique employing layered transmissions and an effective power allocation scheme for the layered transmissions in [29]. Figure 11 shows the sum rate performances of the proposed technique and the existing techniques for the SC-MIMO-NOMA systems
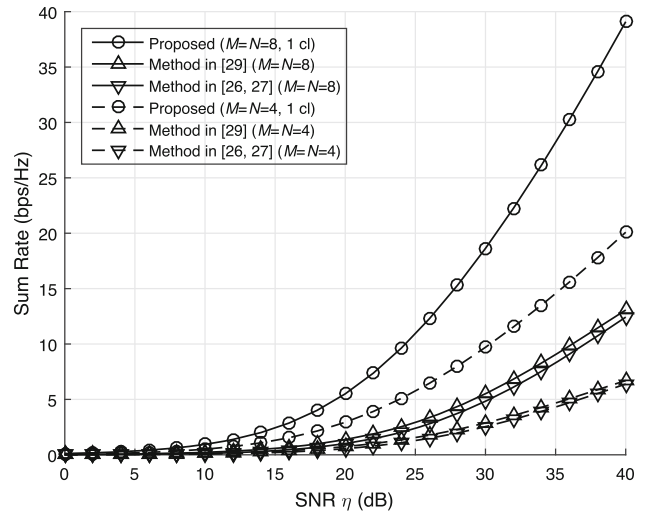

Fig. 11 Comparison of the sum rate performances between the proposed technique and the existing techniques in the SC-MIMO-NOMA systems with $M = N = 4$ and $M = N = 8$
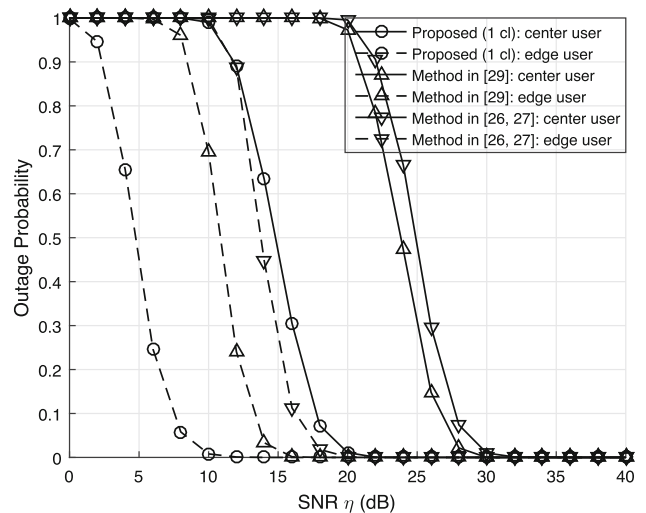

Fig. 12 Comparison of the outage performances for cell-center and cell-edge users between the proposed technique and the existing techniques in the SC-MIMO-NOMA system with $M = N = 4$

with $M = N = 4$ and $M = N = 8$. Figures 12 and 13 show the outage performances of the proposed technique and the existing techniques for the SC-MIMO-NOMA systems with $M = N = 4$ and $M = N = 8$, respectively. The definitions of outage probabilities for cell-center and cell-edge users in the figures are the same as those in Figs. 9 and 10. The existing techniques use time sharing among multiple BSs to avoid MCI.

It can be seen from Fig. 11 that the proposed technique still achieves better sum rate performances than the existing techniques in all the systems. This is because the proposed technique effectively mitigates MCI and optimally allocates the transmit power to data streams to
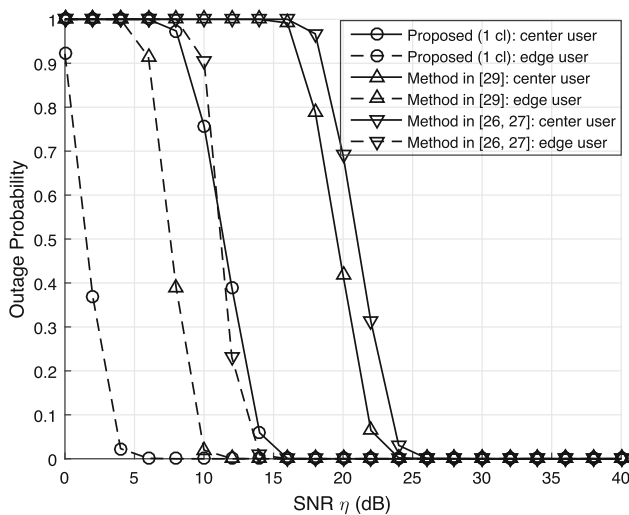
Fig. 13 Comparison of the outage performances for cell-center and cell-edge users between the proposed technique and the existing techniques in the SC-MIMO-NOMA system with $M = N = 8$

maximize the sum rates. In addition, the proposed technique for the SC-MIMO-NOMA systems obtains higher sum rates than the proposed technique for the MC-MIMO-NOMA systems in Figs. 7 and 8 at the expense of fewer clusters simultaneously supported by the serving BS. This is because the CSI combined with the precoder and receive beamformers for ICI elimination tends to be degraded in the MC-MIMO-NOMA systems and the total transmit power is allocated to multiple data streams more effectively in the SC-MIMO-NOMA systems than in the MC-MIMO-NOMA systems. Figures 12 and 13 demonstrate that the proposed technique has lower outage probabilities than the exiting techniques over the entire SNR range. This confirms that the proposed technique can provide the required data rates more reliably than the existing techniques by using effective MCI mitigation and optimal power allocation. Furthermore, the proposed technique for the SC-MIMO-NOMA systems achieves better outage performances than the proposed technique for the MC-MIMO-NOMA systems in Figs. 9 and 10 at the expense of fewer clusters simultaneously supported by the serving BS.

Figures 14 and 15 show the sum rate performances of the proposed technique compared with those of the existing techniques in [23, 24, 26, 27, 29] when the serving cell is interfered from two adjacent cells. The SC-MIMO-NOMA and MC-MIMO-NOMA systems with $M = N = 2$ are considered in Fig. 14, while the systems with $M = N = 4$ are considered in Fig. 15. As expected, the proposed technique achieves better performance than the existing techniques for both SC-MIMO-NOMA and MC-MIMO-NOMA systems, which demonstrates the superiority of the proposed technique. In addition, it is reaffirmed that the proposed technique for the SC-MIMO-NOMA systems
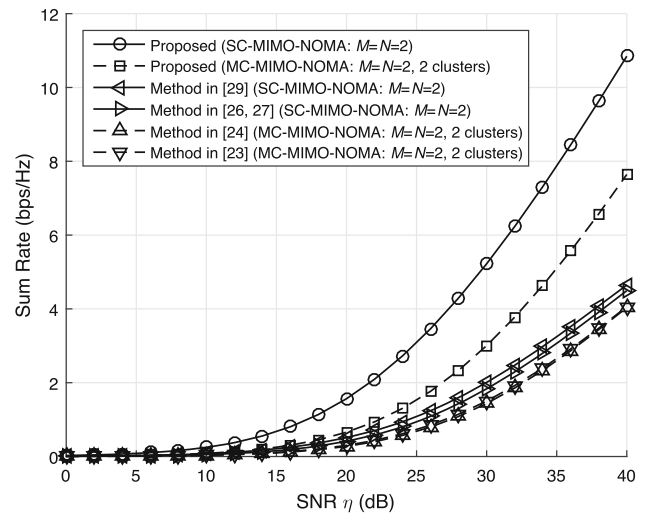

Fig. 14 Comparison of the sum rate performances between the proposed technique and the existing techniques in the SC-MIMO-NOMA and MC-MIMO-NOMA systems with $M = N = 2$ when the serving cell is interfered from two adjacent cells
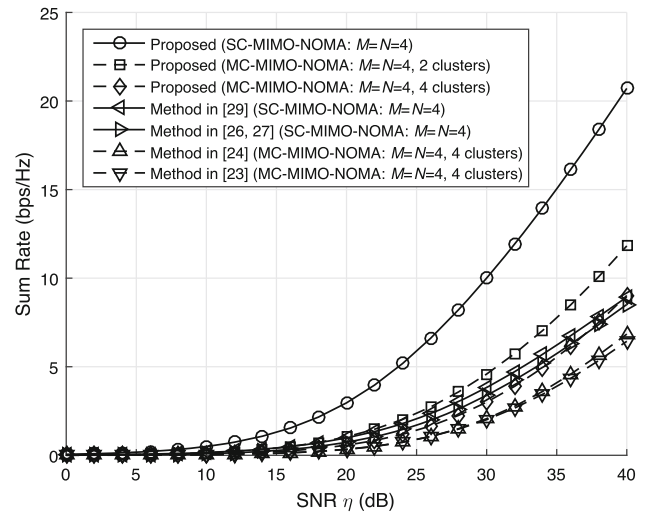

Fig. 15 Comparison of the sum rate performances between the proposed technique and the existing techniques in the SC-MIMO-NOMA and MC-MIMO-NOMA systems with $M = N = 4$ when the serving cell is interfered from two adjacent cells

achieves higher sum rates than the one for the MC-MIMO-NOMA systems at the expense of fewer clusters served simultaneously. This is because the combined CSI tends to be degraded in the MC-MIMO-NOMA systems, while the transmit power is allocated more effectively to multiple data streams in the SC-MIMO-NOMA systems. Compared with the proposed technique in Figs. 7 and 11 for the systems with $M = N = 4$, the proposed technique in Fig. 15 obtains higher sum rates. This is because the number of interfering adjacent cells is reduced from three to two, thereby decreasing MCI.

5 Conclusion

To improve spectral efficiency by transmitting multiple NOMA modulated data streams to each cluster in multi-cell systems, we proposed the MC-MIMO-NOMA technique to design transmit and receive beamforming matrices and perform power allocation between clusters and between data streams in each cluster. In detail, cell-edge users design their own receive beamforming matrices to minimize the MCI power, irrespective of the number of interfering cells. By using the receive beamforming matrices, the serving BS constructs its own transmit beamforming matrices to eliminate ICI between cell-edge users and maximize the lower bounds of the received SNRs at the cell-edge users. Cell-center users then use the transmit beamforming matrices to find their own receive beamforming matrices that eliminate ICI between them and maximize the lower bounds of their received SNRs. Since all users can directly find their own receive beamforming matrices without channel feedback to the serving BS, the proposed technique decreases system overhead. Furthermore, based on the transmit and receive beamforming matrices, the proposed technique allocates the transmit power to clusters and users of each cluster in a manner that maximizes the sum rate performance under power constraints supporting fairness between clusters and between users in each cluster. The simulation results demonstrated that the proposed technique achieves better sum rate and outage performances than the existing techniques and may provide a good balance between the performances and the number of clusters simultaneously supported by the serving BS. This confirms that the proposed technique is more beneficial than the existing techniques. Despite these advantages, the proposed technique has some limitations. Imperfect CSI feedback from cell-edge users may cause performance degradation. The proposed iterative power allocation imposes a computational burden. Addressing these technical issues will be a future research topic.

Data availability Not applicable.

Code availability Not applicable.

References

1. White paper (June 2021). *Ericsson Mobility Report*. https://www.ericsson.com/4a03c2/assets/local/mobility-report/documents/2021/june-2021-ericsson-mobility-report.pdf.

2. Matthaiou, M., Yurduseven, O., Ngo, H. Q., Morales-Jimenez, D., Cotton, S. L., & Fusco, V. F. (2021). The road to 6G: Ten physical layer challenges for communications engineers. *IEEE Communications Magazine, 59*(1), 64–69. https://doi.org/10.1109/MCOM.001.2000208

3. Gui, G., Liu, M., Tang, F., Kato, N., & Adachi, F. (2020). 6G: Opening new horizons for integration of comfort, security, and intelligence. *IEEE Transactions on Wireless Communications, 27*(5), 126–132. https://doi.org/10.1109/MWC.001.1900516

4. Tariq, F., Khandaker, M. R. A., Wong, K.-K., Imran, M. A., Bennis, M., & Debbah, M. (2020). A speculative study on 6G. *IEEE Transactions on Wireless Communications, 27*(4), 118–125. https://doi.org/10.1109/MWC.001.1900488

5. Giordani, M., Polese, M., Mezzavilla, M., Rangan, S., & Zorzi, M. (2020). Toward 6G networks: Use cases and technologies. *IEEE Communications Magazine, 58*(3), 55–61. https://doi.org/10.1109/MCOM.001.1900411

6. Yang, K., Yang, N., Ye, N., Jia, M., Gao, Z., & Fan, R. (2019). Non-orthogonal multiple access: Achieving sustainable future radio access. *IEEE Communications Magazine, 57*(2), 116–121. https://doi.org/10.1109/MCOM.2018.1800179

7. Wang, Q., Zhang, R., Yang, L.-L., & Hanzo, L. (2018). Non-orthogonal multiple access: A unified perspective. *IEEE Wireless Communications, 25*(2), 10–16. https://doi.org/10.1109/MWC.2018.1700070

8. Biswas, S., Bishnu, A., Khan, F. A., & Ratnarajah, T. (2021). In-band full-duplex dynamic spectrum sharing in beyond 5G networks. *IEEE Communications Magazine, 59*(7), 54–60. https://doi.org/10.1109/MCOM.001.2000929

9. da Silva, J. M. B., Wikström, G., Mungara, R. K., & Fischione, C. (2021). Full duplex and dynamic TDD: Pushing the limits of spectrum reuse in multi-cell communications. *IEEE Wireless Communications, 28*(1), 44–50. https://doi.org/10.1109/MWC.001.2000233

10. Papageorgiou, G. K., Voulgaris, K., Ntougias, K., Ntaikos, D. K., Butt, M. M., Galiotto, C., Marchetti, N., Frascolla, V., Annouar, H., Gomes, A., Morgado, A. J., Pesavento, M., Ratnarajah, T., Gopala, K., Kaltenberger, F., Slock, D. T. M., Khan, F. A., & Papadias, C. B. (2020). Advanced dynamic spectrum 5G mobile networks employing licensed shared access. *IEEE Communications Magazine, 58*(7), 21–27. https://doi.org/10.1109/MCOM.001.1900742

11. Chen, S., Sun, S., Xu, G., Su, X., & Cai, Y. (2020). Beam-space multiplexing: Practice, theory, and trends, from 4G TD-LTE, 5G, to 6G and beyond. *IEEE Wireless Communications, 27*(2), 162–172. https://doi.org/10.1109/MWC.001.1900307

12. Liu, W., & Wang, Z. (2019). Non-uniform full-dimension MIMO: New topologies and opportunities. *IEEE Wireless Communications, 26*(2), 124–132. https://doi.org/10.1109/MWC.2019.1800141

13. Sun, Q., Han, S., Chin-Lin, I., & Pan, Z. (2015). On the ergodic capacity of MIMO NOMA systems. *IEEE Wireless Communications Letters, 4*(4), 405–408. https://doi.org/10.1109/LWC.2015.2426709

14. Zeng, M., Yadav, A., Fan, P., Dobre, O. A., Tsiropoulos, G. I., & Poor, H. V. (2017). Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster. *IEEE Journal on Selected Areas in Communications, 35*(10), 2413–2424. https://doi.org/10.1109/JSAC.2017.2725879

15. Yuan, Y., Yuan, Z., & Tian, L. (2020). 5G non-orthogonal multiple access study in 3GPP. *IEEE Communications Magazine, 58*(7), 90–96. https://doi.org/10.1109/MCOM.001.1900450

16. Yuan, Y., Wang, S., Wu, Y., Poor, H. V., Ding, Z., You, X., & Hanzo, L. (2021). NOMA for next-generation massive IoT: Performance potential and technology directions. *IEEE Communications Magazine, 59*(7), 115–121. https://doi.org/10.1109/MCOM.001.2000997

17. Saito, Y., Kishiyama, Y., Benjebbour, A., Nakamura, T., & Higuchi, K. (2013). Non-orthogonal multiple access (NOMA) for cellular future radio access. *IEEE Vehicular Technology Conference (VTC)*. https://doi.org/10.1109/VTCSpring.2013.6692652

18. Higuchi, K., & Kishiyama, Y. (2015). Non-orthogonal multiple access using intra-beam superposition coding and successive interference cancellation for cellular MIMO Downlink. *IEICE Transactions on Communications, E98–B*(9), 1888–1895. https://doi.org/10.1587/transcom.E98.B.1888

19. Chen, Z., Ding, Z., & Dai, X. (2016). Beamforming for combating inter-cluster and intra-cluster interference in hybrid NOMA systems. *IEEE Access, 4*, 4452–4463. https://doi.org/10.1109/ACCESS.2016.2598380

20. Chen, Z., Ding, Z., Dai, X., & Karagiannidis, G. K. (2016). On the application of quasi-degradation to MISO-NOMA downlink. *IEEE Transactions on Signal Processing, 64*(23), 6174–6189. https://doi.org/10.1109/TSP.2016.2603971

21. Chen, C., Cai, W., Cheng, X., Yang, L., & Jin, Y. (2017). Low complexity beamforming and user selection schemes for 5G MIMO-NOMA systems. *IEEE Journal on Selected Areas in Communications, 35*(12), 2708–2722. https://doi.org/10.1109/JSAC.2017.2727229

22. Chen, X., Zhang, Z., Zhong, C., & Ng, D. W. K. (2017). Exploiting multiple-antenna techniques for non-orthogonal multiple access. *IEEE Journal on Selected Areas in Communications, 35*(10), 2207–2220. https://doi.org/10.1109/JSAC.2017.2724420

23. Ding, Z., Adachi, F., & Poor, H. V. (2016). The application of MIMO to non-orthogonal multiple access. *IEEE Transactions on Wireless Communications, 15*(1), 537–552. https://doi.org/10.1109/TWC.2015.2475746

24. Ding, Z., Schober, R., & Poor, H. V. (2016). A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment. *IEEE Transactions on Wireless Communications, 15*(6), 4438–4454. https://doi.org/10.1109/TWC.2016.2542066

25. Shin, W., Vaezi, M., Lee, B., Love, D. J., Lee, J., & Poor, H. V. (2017). Coordinated beamforming for multi-cell MIMO-NOMA. *IEEE Communications Letters, 21*(1), 84–87. https://doi.org/10.1109/LCOMM.2016.2615097

26. Ma, Z., Ding, Z., Fan, P., & Tang, S. (2016). A general framework for MIMO uplink and downlink transmissions in 5G multiple access. *IEEE Vehicular Technology Conference (VTC)*. https://doi.org/10.1109/VTCSpring.2016.7504070

27. Chen, Z., Ding, Z., Dai, X., & Schober, R. (2019). Asymptotic performance analysis of GSVD-NOMA systems with a large-scale antenna array. *IEEE Transactions on Wireless Communications, 18*(1), 575–590. https://doi.org/10.1109/TWC.2018.2883102

28. Paige, C. C., & Saunders, M. A. (1981). Towards a generalized singular value decomposition. *SIAM Journal on Numerical Analysis, 18*(3), 398–405. https://doi.org/10.1137/0718026

29. Choi, J. (2016). On the power allocation for MIMO-NOMA systems with layered transmissions. *IEEE Transactions on Wireless Communications, 15*(5), 3226–3237. https://doi.org/10.1109/TWC.2016.2518182

30. Foschini, G. J. (1996). Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas. *Bell Labs Technical Journal, 1*(2), 41–59. https://doi.org/10.1002/bltj.2015

31. Wolniansky, P. W., Foschini, G. J., Golden, G. D., & Valenzuela, R. (1998). V-BLAST: An architecture for realizing very high data rates over the rich-scattering wireless channel. In *URSI International symposium on signals, systems, and electronics (ISSSE)* (pp. 295–300). https://doi.org/10.1109/ISSSE.1998.738086

32. Jiang, M., Li, Y., Zhang, Q., Li, Q., & Qin, J. (2018). MIMO beamforming design in nonorthogonal multiple access downlink interference channels. *IEEE Transactions on Vehicular Technology, 67*(8), 6951–6959. https://doi.org/10.1109/TVT.2018.2821166

33. Oviedo, J. A., & Sadjadpour, H. R. (2017). A fair power allocation approach to NOMA in multiuser SISO systems. *IEEE Transactions on Vehicular Technology, 66*(9), 7974–7985. https://doi.org/10.1109/TVT.2017.2689000

Changyong Shin received the B.S. and M.S. degrees from Yonsei University, Seoul, South Korea in 1993 and 1995, respectively, and the Ph.D. degree from the University of Texas at Austin, in 2006, all in electrical engineering. From 1995 to 2001, he was a Senior Research Engineer at LG Electronics Inc., Seoul, South Korea, where he worked on digital video signal processing and VLSI circuit design for digital signal processing. From 2007 to 2014, he was a Principal Research Engineer at Samsung Advanced Institute of Technology (SAIT), Gyunggi, South Korea, where he worked on future wireless communication systems including next-generation cellular systems. Since 2014, he has been with the Department of Smart Information and Communications Engineering at Sun Moon University, Chungnam, South Korea. His research interests include wireless communications and signal processing for communications.