CrossMark

Non-orthogonal multiple access with joint maximum likelihood detection in heterogeneous network

Yukitoshi Sanada

Abstract

As one of the key technologies in the fifth-generation mobile communication system, non-orthogonal multiple access (NOMA) has been investigated. In NOMA, multiple terminals are assigned the same frequency resources by a scheduler on the basis of the difference in propagation losses between a base station and user terminals. Each terminal cancels the signals for the other terminals and extracts its desired signal. On the other hand, the application of joint maximum likelihood (ML) detection to overloaded signals has also been investigated, and joint ML detection can be applied to a NOMA downlink. In this paper, the effect of joint ML detection in a heterogeneous NOMA network is presented. The numerical results obtained through system-level simulation show that joint ML detection in a heterogeneous NOMA downlink can effectively offload mobile traffic from a macro base station to a pico base station. It is shown that a heterogeneous NOMA network with joint ML detection improves the throughput performance by 0.2 bit/user/subcarrier as compared to that without joint ML detection at a cumulative probability of 0.5. The system throughput is also increased about twofold with joint ML detection.

Keywords: Heterogeneous network, Proportional fairness scheduling, Joint ML detection

1 Introduction

The amount of mobile traffic has increased explosively in recent years owing to the rise in the penetration rate of mobile devices such as smart phones. It is expected to increase 1000-fold from 2010 to 2020 [1]. To accommodate such a large amount of mobile traffic, the concept of the fifth-generation mobile communication system (5G) has been discussed.

As one of the key technologies in 5G, non-orthogonal multiple access (NOMA) has been proposed and investigated as a mean to improve the spectrum efficiency of the system [2–7]. In the downlink of the NOMA scheme, multiple user terminals are assigned the same frequency resources. These user terminals are located at different positions relative to a base station, and their propagation losses are different. The base station transmits signals to these user terminals with different transmission powers. The receiver in the user terminal that is closer to the base station eliminates the signals for the other user terminals

by interference cancelation and extracts the desired signal. Under practical modulation and coding parameters that enable limited decoding and interference cancelation capability, the powers of the received signals must be sufficiently different to demodulate and decode the undesired signals through interference cancelation. Otherwise, the residual interference caused by unreliable replica symbols will deteriorate the demodulation performance of the desired signal [8]. The improvement in throughput by the NOMA scheme is limited owing to the residual interference even though the number of terminals assigned the same resource is set to more than two [9].

On the other hand, in the LTE-Advanced system, a heterogeneous network (HetNet) has been specified in the standard [10–12]. In HetNets, small cells called pico cells are placed within a macro cell, and each user terminal is connected to a nearby pico cell to prevent traffic congestion in a macro cell network. However, the interference owing to the signal from the macro base station (MBS) limits the traffic offloading capability of HetNets. The transmission power of the MBS is much larger than that

Correspondence: sanada@elec.keio.ac.jp
Department of Electronics and Electrical Engineering, Keio University, 3-14-1 Hiyoshi, Kohoku, Yokohama 2238522, Japan

of the pico base station (PBS), and the signal from the PBS must be protected [13, 14].

Some recent studies have treated this problem with NOMA schemes in the downlink [15–17]. However, not many studies have taken demodulation schemes into account. Codeword-level successive interference cancelation (SIC) requires turbo decoding and replica signal generation. These processes require coding parameters as well as modulation parameters of the signals for the other users. They also cause a demodulation delay that is not acceptable in some applications. Resource block assignment among NOMA users should coincide to reduce the complexity of SIC. If the desired signal with a smaller signal power overlaps with multiple undesired signals, a receiver has to decode all these interference signals for the generation of SIC replicas.

If the number of antenna elements implemented in a user terminal is more than the total number of the desired and interference signals, it is possible to separate the desired signal and the other interfering signals by a normal multiple-input multiple-output (MIMO) detection process. In this case, simple linear detection schemes such as a zero-forcing detection scheme or a minimum mean-square-error detection scheme can be applied. However, owing to the limitation of its form factor, the number of receive antennas may be less than the number of received signal streams. This is an "overloaded" situation, and a nonlinear detection scheme must be applied in the receiver. The joint detection of multiple signal streams in the receiver is a solution to this limitation although joint detection increases the demodulation complexity [18]. Joint detection only requires the channel responses and modulation orders of the interfering signals. In [19], the effect of joint detection in terminals both near and far from a base station located at the center of a cell was investigated. In this case, joint detection works effectively in cell site terminals. In a heterogeneous network, joint detection works effectively in the cell site terminals of a pico cell as they suffer from large interference from a MBS.

Thus, in this paper, the application of joint maximum likelihood (ML) detection to a heterogeneous NOMA network is investigated. The numerical results obtained through system-level simulation show that joint ML detection in a NOMA downlink alleviates the effect of the interference and can offload mobile traffic from the MBS to the PBS.

This manuscript is organized as follows. Section 2 presents the application of the joint ML detection scheme in a heterogeneous NOMA network. The numerical results of the proposed system obtained through system-level simulation are presented in Section 3. Finally, Section 4 gives our conclusions.

2 System model

2.1 Signal model

In this study, the joint ML detection of overloaded signals is applied to a heterogeneous NOMA network. It is assumed here that a MBS and a PBS are operated in the same frequency band. A PBS exists in each sector of the macro cell and offloads the mobile traffic from the MBS as shown in Fig. 1. At the receiver, joint ML detection is applied to demodulate the desired signal.

Here, it is assumed that a macro user (MU) receives the signal from the MBS, and a pico user (PU) receives the signal from the PBS. User assignment is carried out in each resource block, allowing the same user to be a MU and a PU in different resource blocks. It is also assumed that the PBS is under the control of the MBS in terms of the user assignment, and NOMA between a PU and a MU is assumed.

Suppose that U_m is an index for the MU, and U_p is an index for the PU. It is then assumed that $x^l_{\mathrm{MU}_m}(k_m)$ is the transmitted symbol on the lth subcarrier for the MU from the MBS, $x^l_{\mathrm{PU}_p}(k_p)$ is the transmitted symbol on the lth subcarrier for the PU from the PBS, and the numbers of constellation points in one symbol are $N_m(0 \leq k_m \leq N_m - 1)$ and $N_p(0 \leq k_p \leq N_p - 1)$ in the signals for the MU and PU, respectively. The channel responses and
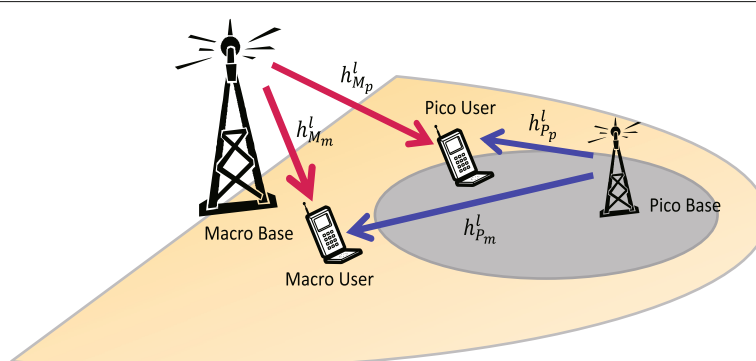

Fig. 1 HetNet channel model

their phase rotation terms from the MBS to the MU, from the MBS to the PU, from the PBS to the MU, and from the PBS to the PU on the lth subcarrier are defined as $h^l_{\mathrm{MU}_m}$ and $\exp\left(j\theta^l_{\mathrm{MU}_m}\right)$, $h^l_{\mathrm{MU}_p}$ and $\exp\left(j\theta^l_{\mathrm{MU}_p}\right)$, $h^l_{\mathrm{PU}_m}$ and $\exp\left(j\theta^l_{\mathrm{PU}_m}\right)$, and $h^l_{\mathrm{PU}_p}$ and $\exp\left(j\theta^l_{\mathrm{PU}_p}\right)$, respectively, and they are given as follows:

$$h^l_{\mathrm{MU}_m} = \left|h^l_{\mathrm{MU}_m}\right| \exp\left(j\theta^l_{\mathrm{MU}_m}\right), \tag{1}$$

$$h^l_{\mathrm{MU}_p} = \left|h^l_{\mathrm{MU}_p}\right| \exp\left(j\theta^l_{\mathrm{MU}_p}\right), \tag{2}$$

$$h^l_{\mathrm{PU}_m} = \left|h^l_{\mathrm{PU}_m}\right| \exp\left(j\theta^l_{\mathrm{PU}_m}\right), \tag{3}$$

$$h^l_{\mathrm{PU}_p} = \left|h^l_{\mathrm{PU}_p}\right| \exp\left(j\theta^l_{\mathrm{PU}_p}\right). \tag{4}$$

The MBS uses power control to reduce the interference to the pico cell while the pico base station transmits the signal with fixed power. The signal received by the U_mth MU is then given as:

$$y^l_m = \alpha^l_M h^l_{\mathrm{MU}_m} x^l_{\mathrm{MU}_m}(k_m) + h^l_{\mathrm{PU}_m} x^l_{\mathrm{PU}_p}(k_p) + z^l_m, \tag{5}$$

where α^l_M is the power control coefficient, which takes a value of less than 1.0, and z^l_m is additive white Gaussian noise (AWGN) with a mean of 0 and a variance of σ^2. On the other hand, the signal received by the U_pth PU is:

$$y^l_p = h^l_{\mathrm{PU}_p} x^l_{\mathrm{PU}_p}(k_p) + \alpha^l_M h^l_{M_{U_p}} x^l_{\mathrm{MU}_m}(k_m) + z^l_p, \tag{6}$$

where z^l_p is also AWGN with a mean of 0 and a variance of σ^2.

2.2 Throughput of NOMA with joint ML detection in HetNet

The constellation of the superimposed signals at the receiver changes with the difference in the signal phases. The difference in the signal phases at the MU is given as:

$$\theta^l_{U_m} = \left|\theta^l_{\mathrm{MU}_m} - \theta^l_{\mathrm{PU}_m}\right|, \tag{7}$$

while at the PU it is given as:

$$\theta^l_{U_p} = \left|\theta^l_{\mathrm{MU}_p} - \theta^l_{\mathrm{PU}_p}\right|. \tag{8}$$

In this study, as explained in detail in the Appendix, the constellation constraint capacity is calculated as the throughput of the system [20–22]. Furthermore, it is assumed that α^l_M, N_m, and N_p are determined in each resource block.

The throughput of a MU that is without joint ML detection and demodulates only the desired signal is given as:

$$T_{cm}\left(U_m, l | U_p, h^l_{\mathrm{MU}_m}, h^l_{\mathrm{PU}_m}, \alpha^l_M, N_m, N_p\right)$$

$$= \log_2(N_m) - \frac{1}{N_m N_p} \sum_{k_m=0}^{N_m-1} \sum_{k_p=0}^{N_p-1} \mathop{E}_{\theta^l_{U_m}, z^l_m}$$

$$\left[\log_2\left(\sum_{i_m=0}^{N_m-1} \exp\left(-\left|\alpha^l_M h^l_{\mathrm{MU}_m}\right| \exp\left(j\theta^l_{U_m}\right)\left(x^l_{\mathrm{MU}_m}(k_m) - x^l_{\mathrm{MU}_m}(i_m)\right)\right.\right.\right.$$

$$\left.\left. + \left|h^l_{\mathrm{PU}_m}\right| x^l_{\mathrm{PU}_p}(k_p) + z^l_m\right|^2 / \sigma^2\right)$$

$$\left. - \log_2\left(\exp\left(-\left|\left|h^l_{\mathrm{PU}_m}\right| x^l_{\mathrm{PU}_p}(k_p) + z^l_m\right|^2 / \sigma^2\right)\right)\right]. \tag{9}$$

In the same way as for the MU, the throughput of a PU that demodulates only the desired signal is calculated as:

$$T_{cp}\left(U_p, l | U_m, h^l_{\mathrm{PU}_p}, h^l_{\mathrm{PU}_m}, \alpha^l_M, N_m, N_p\right)$$

$$= \log_2(N_p) - \frac{1}{N_m N_p} \sum_{k_m=0}^{N_m-1} \sum_{k_p=0}^{N_p-1} \mathop{E}_{\theta^l_{U_p}, z^l_p}$$

$$\left[\log_2\left(\sum_{i_p=0}^{N_p-1} \exp\left(-\left|\left|h^l_{\mathrm{PU}_p}\right| \exp\left(j\theta^l_{U_p}\right)\left(x^l_{\mathrm{PU}_p}(k_p) - x^l_{\mathrm{PU}_p}(i_p)\right)\right.\right.\right.\right.$$

$$\left.\left. + \alpha^l_M \left|h^l_{\mathrm{MU}_p}\right| x^l_{\mathrm{MU}_m}(k_m) + z^l_p\right|^2 / \sigma^2\right)$$

$$\left. - \log_2\left(-\left|\left|h^l_{\mathrm{MU}_p}\right| x^l_{\mathrm{MU}_m}(k_m) + z^l_p\right|^2 / \sigma^2\right)\right)\right]. \tag{10}$$

On the other hand, in the assumed NOMA heterogeneous network, joint ML detection is applied at the receivers of the MU and the PU. Suppose that $T_m(U_p, l)$ and $T_m(U_p, l)$ are the instantaneous throughputs of the U_mth MU and U_pth PU on the lth subcarrier, respectively. The throughput of the MU is calculated as Eq. (11) and that of the PU is calculated as Eq. (12).

$$T_m\left(U_m, l | U_p, h^l_{\mathrm{MU}_m}, h^l_{\mathrm{PU}_m}, \alpha^l_M, N_m, N_p\right)$$

$$= \log_2(N_m) - \frac{1}{N_m N_p} \sum_{k_m=0}^{N_m-1} \sum_{k_p=0}^{N_p-1} \mathop{E}_{\theta^l_{U_m}, z^l_m}$$

$$\left[\log_2\left(\sum_{i_m=0}^{N_m-1} \sum_{i_p=0}^{N_p-1} \exp\left(-\left|\alpha^l_M\right| h^l_{\mathrm{MU}_m}\left| \exp\left(j\theta^l_{U_m}\right)\left(x^l_{\mathrm{MU}_m}(k_m) - x^l_{\mathrm{MU}_m}(i_m)\right)\right.\right.\right.\right.$$

$$\left.\left. + \left|h^l_{\mathrm{PU}_m}\right|\left(x^l_{\mathrm{PU}_p}(k_p) - x^l_{\mathrm{PU}_p}(i_p)\right) + z^l_m\right|^2 / \sigma^2\right)$$

$$\left. - \log_2\left(\sum_{i_p=0}^{N_p-1} \exp\left(-\left|\left|h^l_{\mathrm{PU}_m}\right|\left(x^l_{\mathrm{PU}_p}(k_p) - x^l_{\mathrm{PU}_p}(i_p)\right) + z^l_m\right|^2 / \sigma^2\right)\right)\right] \tag{11}$$

and

$$T_p\left(U_p, l|U_m, h^l_{\mathrm{PU}_p}, h^l_{\mathrm{MU}_p}, \alpha^l_M, N_m, N_p\right)$$

$$= \log_2(N_p) - \frac{1}{N_m N_p} \sum_{k_m=0}^{N_m-1} \sum_{k_p=0}^{N_p-1} \underset{\theta^l_{U_p}, z^l_p}{E}$$

$$\left[\log_2\left(\sum_{i_m=0}^{N_m-1} \sum_{i_p=0}^{N_p-1} \exp\left(- \left|\left| h^l_{\mathrm{PU}_p} \right| \exp\left(j\theta^l_{U_p}\right) \left(x^l_{\mathrm{PU}_p}(k_p) - x^l_{\mathrm{PU}_p}(i_p)\right) \right.\right.\right.$$

$$\left.\left.\left. -\alpha^l_M \left| h^l_{\mathrm{MU}_m} \right| \left(x^l_{\mathrm{MU}_m}(k_m) - x^l_{\mathrm{MU}_m}(i_m)\right) + z^l_p \right|^2 / \sigma^2 \right)\right.$$

$$\left. -\log_2\left(\sum_{i_m=0}^{N_m-1} \exp\left(- \left|\alpha^l_M \left| h^l_{\mathrm{MU}_p}\right| \left(x^l_{\mathrm{MU}_m}(k_m) - x^l_{\mathrm{MU}_m}(i_m)\right) + z^l_p\right|^2 / \sigma^2\right)\right) \right].$$

$$(12)$$

2.3 Proportional fairness scheduling

For user assignment, proportional fairness (PF) scheduling is applied, and all the resource blocks are assigned to the pair of a MU and a PU that achieve the largest PF metric. The PF metric is calculated from the instantaneous throughput and the average throughput as:

$$T_{\mathrm{Ave}}(u, t+1) = \left(1 - \frac{1}{t_c}\right) T_{\mathrm{Ave}}(u, t)$$

$$+ \frac{1}{t_c}\left(\sum_{l_m \in \{\Omega_m\}} T_m\left(u, l_m | U_p, h^{l_m}_u, h^{l_m}_{U_p}, \alpha^{l_m}_M, N_m, N_p\right) \right.$$

$$\left. + \sum_{l_p \in \{\Omega_p\}} T_p\left(u, l_p \left| U_m, h^{l_p}_{U_m}, h^{l_p}_u, \alpha^{l_p}_M, N_m, N_p\right.\right)\right) \quad (13)$$

where u is the user index; $\{\Omega_m\}$ and $\{\Omega_p\}$ are the groups of subcarriers assigned to the uth user as the MU and the PU, respectively; t is the time index; and t_c is the period of the moving average [23]. From the above equation, the PF metric is given as:

$$P_{\mathrm{Prob}}(U_m, U_p) = \underset{U_m, U_p, \alpha^l_M, N_m, N_p}{max} \sum_{l \in \{\Omega_b\}}$$

$$\frac{T_m\left(U_m, l \left| U_p, h^l_{\mathrm{MU}_m}, h^l_{\mathrm{PU}_p}, \alpha^l_M, N_m, N_p\right.\right)}{T_{\mathrm{Ave}}(U_m, t)} \quad (14)$$

$$+ \frac{T_p\left(U_p, l | U_m, |h^l_{\mathrm{PU}_p}, h^l_{\mathrm{MU}_m}, \alpha^l_M, N_m, N_p\right)}{T_{\mathrm{Ave}}(U_p, t)}$$

where $\{\Omega_b\}$ is the set of subcarrier indexes in the bth resource block. Through this user assignment process, the pair of users, the modulation parameters, and the transmission power from the base station that maximizes Eq. (14) are obtained. The user throughput for the uth user is then given as:

$$T(u) = \sum_{l_m \in \{\Omega_m\}} T_m\left(u, l_m | U_p, h^{l_m}_u, h^{l_m}_{U_p}, \alpha^{l_m}_M, N_m, N_p\right)$$

$$+ \sum_{l_p \in \{\Omega_p\}} T_p\left(u, l_p | U_m, h^{l_p}_{U_m}, h^{l_p}_u, \alpha^{l_p}_M, N_m, N_p\right))$$

$$(15)$$

and the user fairness for the uth user is calculated as:

$$F = \sqrt[u]{\prod_{u=1}^{U} T(u)}. \quad (16)$$

3 Method

The throughput performance of the heterogeneous NOMA network with joint ML detection is obtained through system-level simulation. The simulation conditions are presented in Table 1. The throughput is calculated with the parameters such as the signal-to-interference-plus-noise ratio (SINR), the difference in the phases of the received signal streams, and the modulation orders of the transmit symbols that are selected from QPSK, 16QAM, 64QAM, and 256QAM. As the cell layout, the 19-hexagonal-cell site shown in Fig. 2 is assumed, and the minimum distances from a user terminal to the MBS and the PBS are assumed to be 35 m and 10 m, respectively. The three-sector cell model presented in

Table 1 Simulation conditions

|	
Scheduling algorithm	PF scheduling
Modulation scheme	QPSK, 16QAM, 64QAM, 256QAM
Cell layout	19-hexagonal-cell site
Inter-site distance	500 m
Minimum distance (MBS - user)	35 m
Minimum distance (PBS - user)	10 m
Number of user terminals per sector	5, 10, 20
Distribution of user terminals	Uniform
MBS height	35 m
PBS height	10 m
User height	1.5 m
MBS maximum transmission power	43 dBm
PBS total transmission power	30 dBm
Distance-dependent path loss (MBS)	128.1 + 37.6log10(R) dB, R in km
Distance-dependent path loss (PBS)	140.7 + 36.7log10(R) dB R in km
Shadowing standard deviation (MBS)	5 dB
Shadowing standard deviation (PBS)	7 dB
Shadowing correlation	0.5
Channel model	Six-path Rayleigh
System bandwidth	4.32 MHz
Number of resource blocks	24
Resource block bandwidth	180 kHz
Receiver noise density	− 174 dBm/Hz
Time interval	100
User drops	≥ 50
Trial per user drops	30
Number of symbols per trial	100
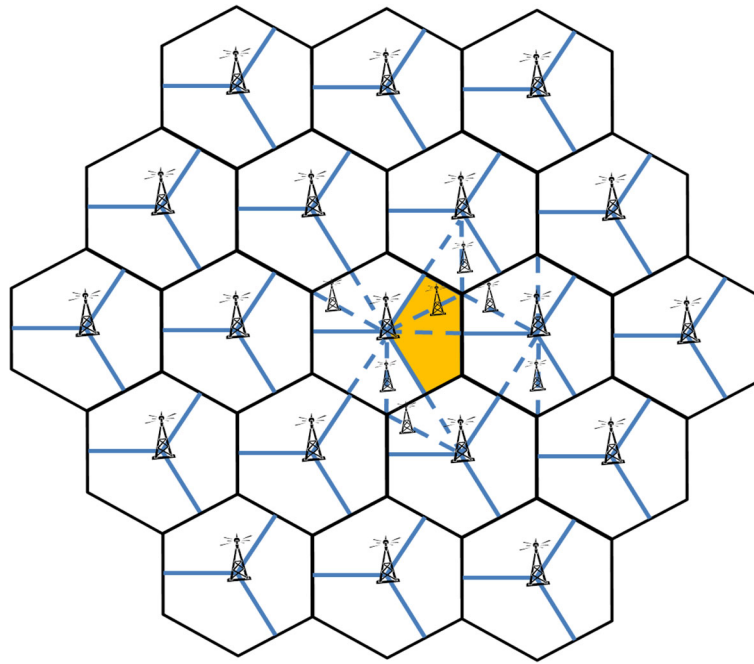
Fig. 2 19-hexagonal-cell site model

Fig. 3 with one PBS in each sector is assumed. The distance between the MBSs is set to 500 m. User terminals are uniformly distributed, and 5, 10, or 20 terminals per sector are assumed. The height of the MBS is set to 35 m and that of the PBS is set to 10 m, and the height of the

Fig. 3 Three-sector cell model

user terminal is set to 1.5 m. The maximum transmission power of the MBS is set to 43 dBm unless otherwise specified, and the transmission power is controlled by the coefficient α_M^l, which is determined through the PF scheduling. On the other hand, the transmission power of the PBS is fixed to 30 dBm. The propagation loss is assumed to be proportional to the traveling distance, and the standard deviation of the shadowing is 5 dB between the MBS and the user terminal and 7 dB between the PBS and the user terminal. The correlation of the shadowing is set to 0.5. The RMS delay spread is 1 μs, the maximum Doppler shift, f_D, is set to 5.55 Hz, and a six-path exponential delay profile fading channel model is assumed. There are 12 subcarriers in 1 resource block, 1 subband consists of 24 resource blocks, 1 resource block occupies 180 kHz, and the frequency bandwidth of the channel is 4.32 MHz. The spectrum density of the noise in the receiver is assumed to be -174 dBm/Hz. In the calculation of the SINR, the interference from the 18 surrounding MBSs and that from the pico cells located in the adjacent sectors to the sector of concern are taken into account. When the PBS is located at the cell edge, it is assumed that a directional antenna is used and no intercell interference from the PBSs in the adjacent cells occurs [24].

The average throughput is calculated over 100 time slots for each user drop. The number of user drops is more than 50 for each condition, and the number of trials per user drop is 30. The number of symbols transmitted per trial
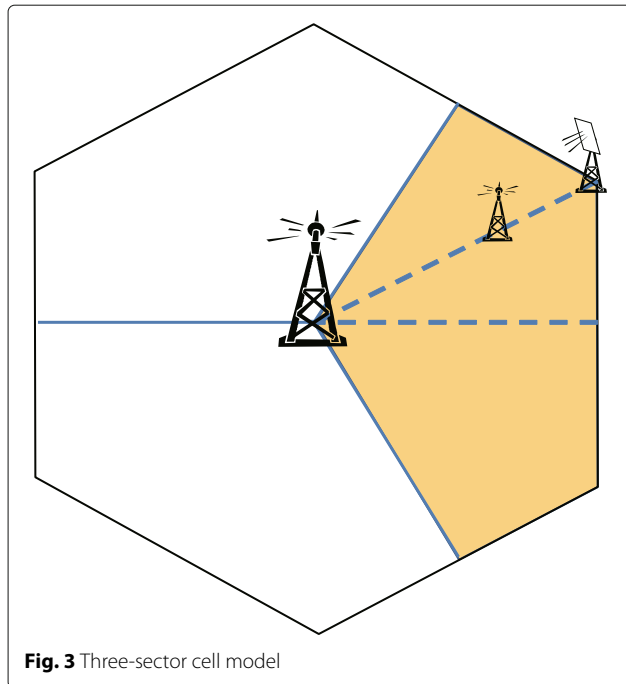
is 100, and the last 80 symbols are used for throughput evaluation.

4 Result and discussion

4.1 Offloading capability

The cumulative distribution functions of user throughputs with and without joint ML detection are presented in Fig. 4. The PBS is located at a distance of two thirds of the cell radius from the cell center, and the number of users per sector is 10. The maximum transmission power of the MBS is 43 dBm. It is clear from the figure that the throughput is improved with joint ML detection. At a cumulative probability of 0.5, the throughput improves by about 0.2 bit/subcarrier/user. The throughputs of the MUs are shown in Fig. 5 while those of the PUs are shown in Fig. 6. From these figures, in the pico cell, the throughputs of the users without joint ML detection are limited owing to the interference from the macro cell. With joint ML detection, the low-throughput users in the macro cell switch to the pico cell and their throughputs improve. At the same time, the high-throughput users in the macro cell also achieve better throughput performance. On the other hand, the numbers of low- and medium-throughput users increase in the pico cell. Thus, joint ML detection offloads mobile traffic and improves the total throughput of the heterogeneous NOMA network.

4.2 Pico base station location

The system throughputs and user fairness curves of the NOMA HetNet versus the PBS location are presented in Figs. 7 and 8, respectively. The number of users per sector is 10. The maximum transmission power of the MBS

is 43 dBm. These figures show that joint ML detection improves the performance regardless of the location of the PBS.

This is because the interference from the MBS is alleviated owing to the joint ML detection in the user terminals. If the PBS is located close to the MBS, it cannot effectively cover the user terminals at the cell edge. If it is too close to the cell edge, the PBS suffers from intercell interference. Thus, the maximum throughput is realized when the PBS is located at a distance of two thirds of the cell radius. When the PBS is located at the cell edge, its coverage area reduces to one third owing to the use of the directional antenna. Thus, without joint ML detection, the throughput and the user fairness are less than those at a distance of five sixths of the cell radius even though no intercell interference from the PBSs in the adjacent cells is assumed. On the other hand, with joint ML detection, the user fairness is greater than that at a distance of five sixths of the cell radius although the throughput is slightly lower. This is because joint ML detection enlarges the coverage area of the PBS, and the PBS realizes better connections to the cell edge users that suffer from intercell interference from the MBSs in the adjacent cells.

The system throughputs and user fairness curves for the case of five user terminals per sector are presented in Figs. 9 and 10, respectively. The same tendencies as those in Figs. 7 and 8 can be observed. When the number of the users is five, the dependence on the PBS location is stronger. This is because the PUs suffer more significantly from the interference signals transmitted for the MUs as the probability of selecting a MU that causes less interference decreases as the number of users decreases.
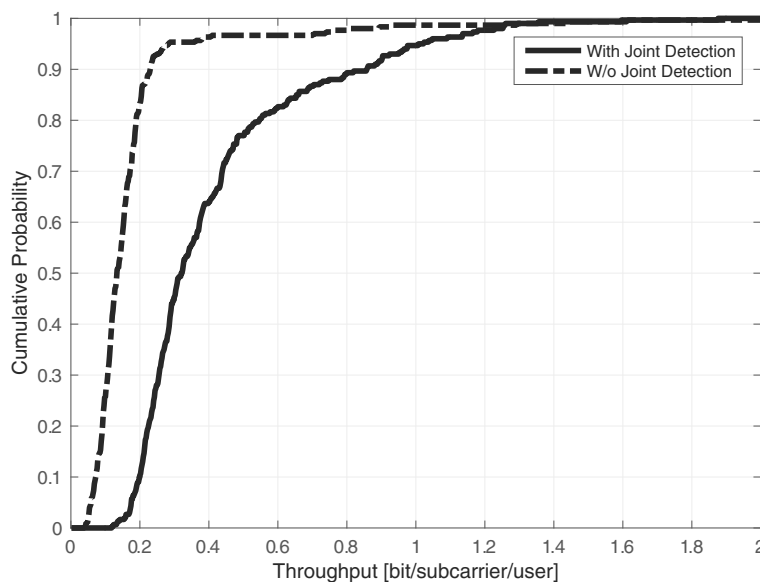

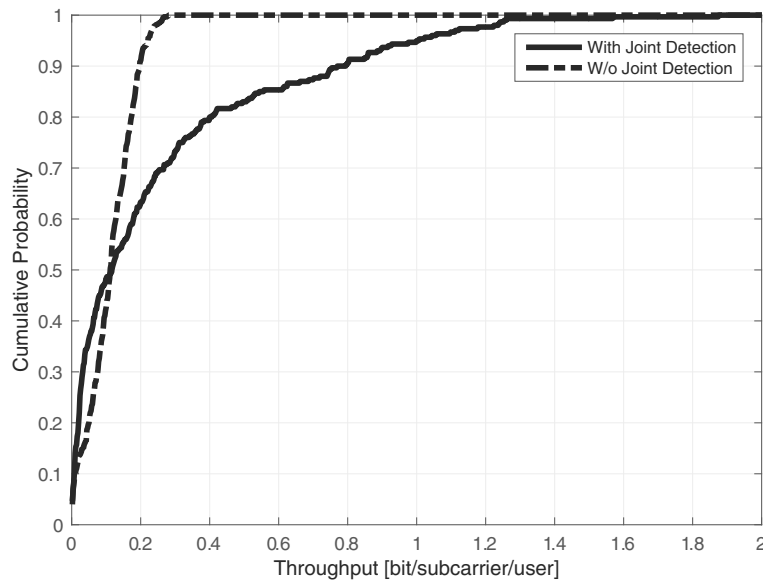**Fig. 4** Total system throughput performance (10 users/sector)

Fig. 5 System throughput performance in macro cell (10 users/sector)

The probability distribution functions of the transmission power of the MBS with the PBS at a distance of two thirds of the cell radius for the cases of 5, 10, and 20 users per sector are shown in Figs. 11, 12, and 13, respectively, where the maximum transmission power of the MBS is 43 dBm. According to the figures, the MBS can transmit a signal with a larger power when the user terminals apply joint ML detection. The base station transmits the signal with its maximum power almost all the time when joint ML detection is applied. On the other hand, without applying joint ML detection, the MBS reduces the transmission power with a larger probability. This is because a larger transmission power from the MBS results in a larger SINR to the user terminals at the edge of the cell. The MBS transmits with a larger power if joint ML detection mitigates the interference to the PU terminals.
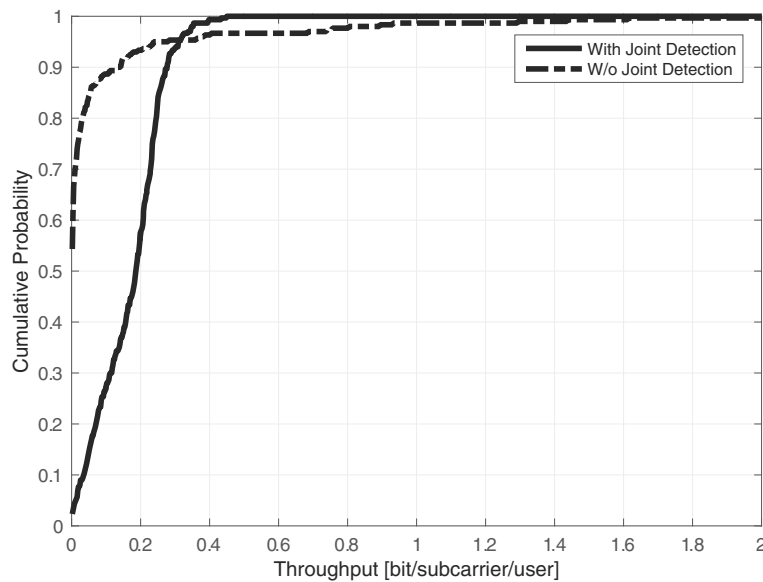

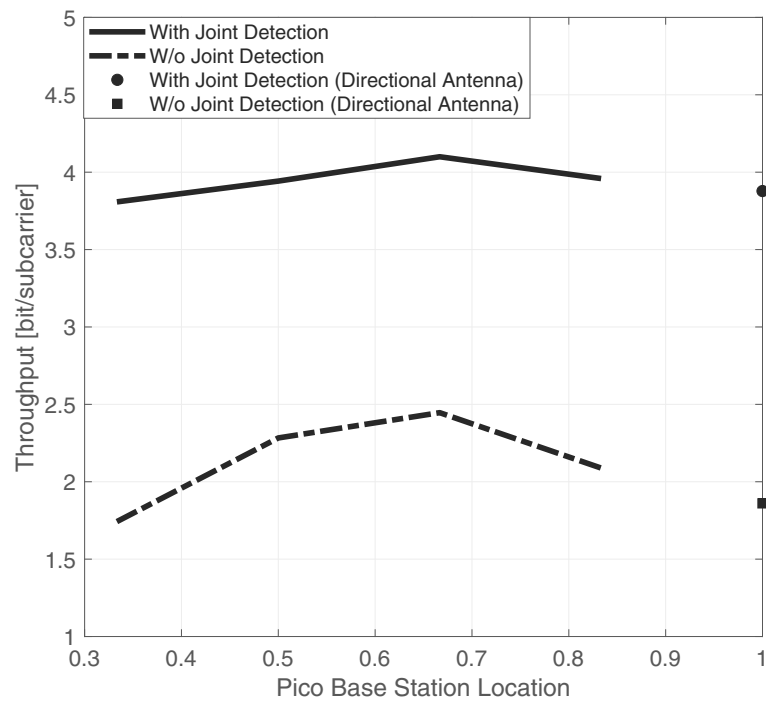**Fig. 6** System throughput performance in pico cell (10 users/sector)

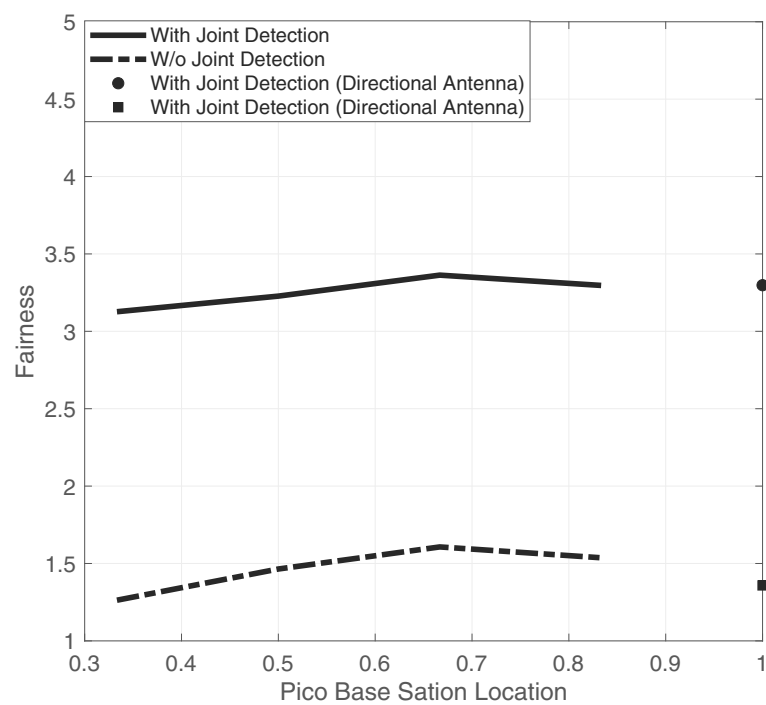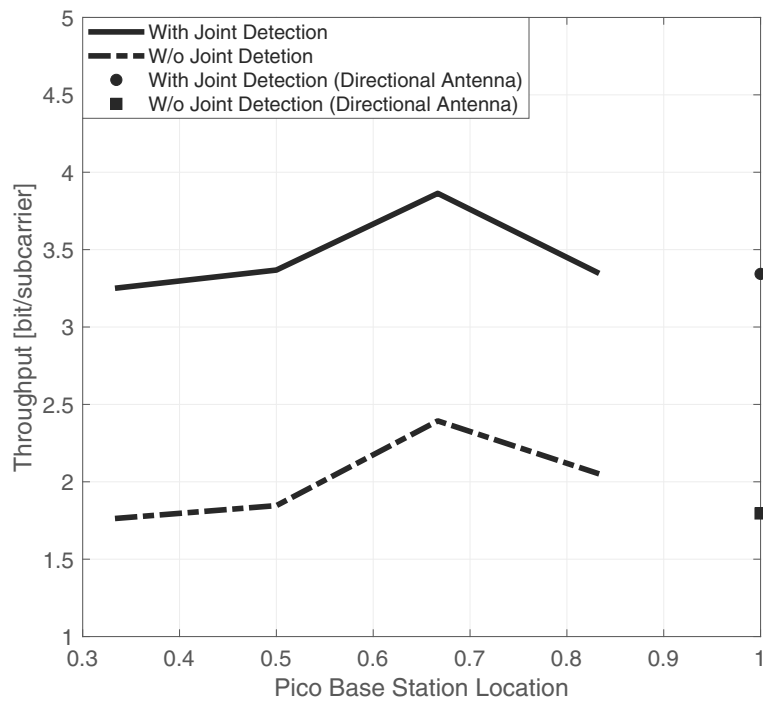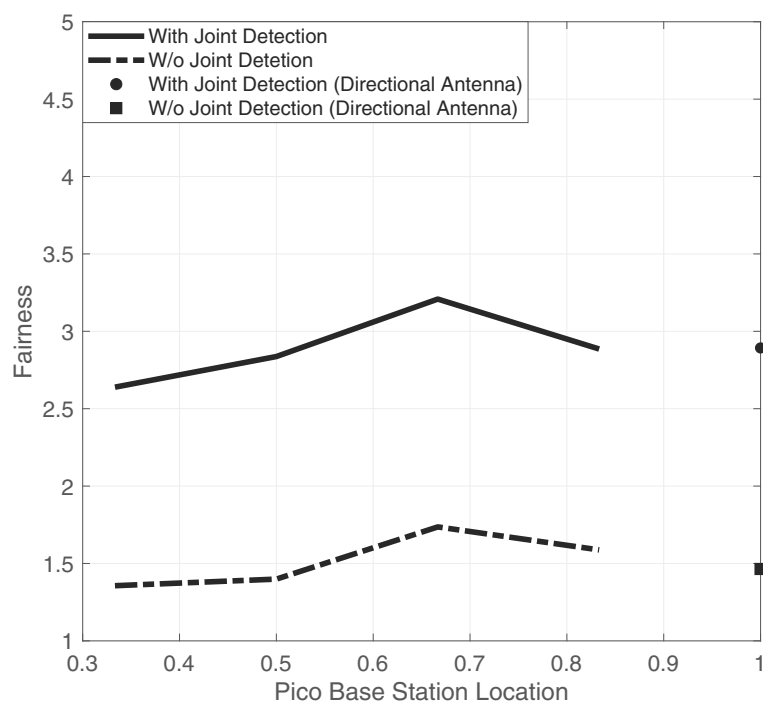**Fig. 7** System throughput versus PBS location (10 users/sector)

Fig. 8 User fairness versus PBS location (10 users/sector)

Fig. 9 System throughput versus PBS location (5 users/sector)

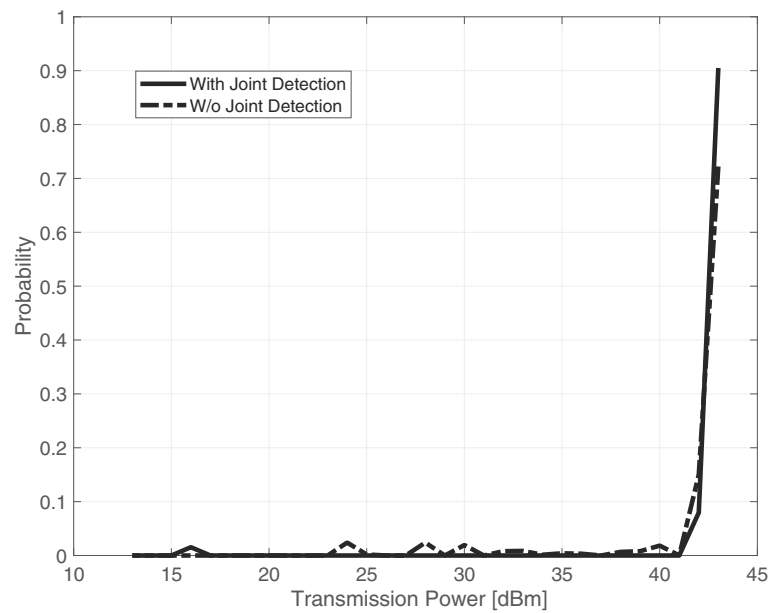**Fig. 10** User fairness versus PBS location (5 users/sector)

Fig. 11 Probability distribution of MBS transmission power (5 users/sector, PBS: 2(cell radius)/3)

5   Conclusions

In this study, a joint ML detection scheme for the demodulation of overloaded signals has been applied to the heterogeneous NOMA network. The joint ML detection effectively alleviates the effect of the interference between the signals from the MBS and PBS. The numerical results obtained through system-level simulation have shown that joint ML detection in the NOMA downlink effectively offloads mobile traffic from the MBS to the PBS. As a result, user fairness improves, and the system throughput increases about twofold. The maximum system throughput is achieved when the PBS is at a distance of two thirds of the cell radius from the base station if joint ML detection is applied.
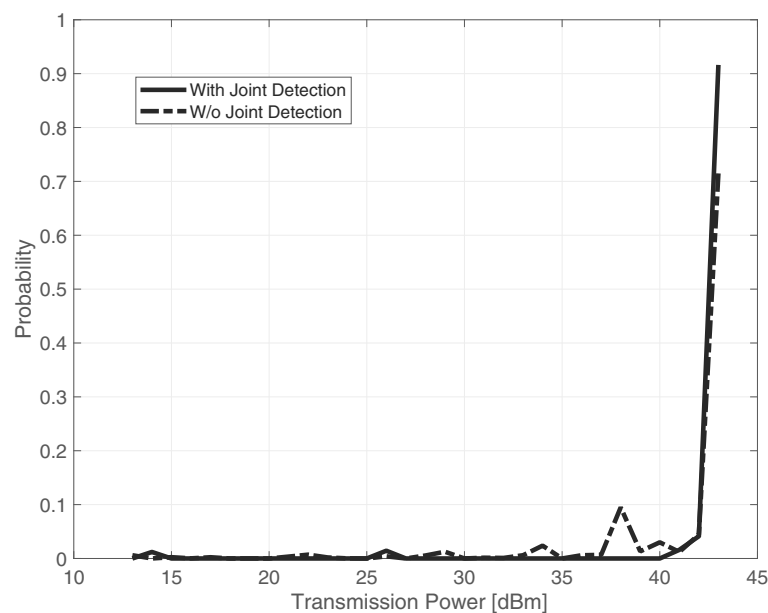

Fig. 12 Probability distribution of MBS transmission power (10 users/sector, PBS: 2(cell radius)/3)
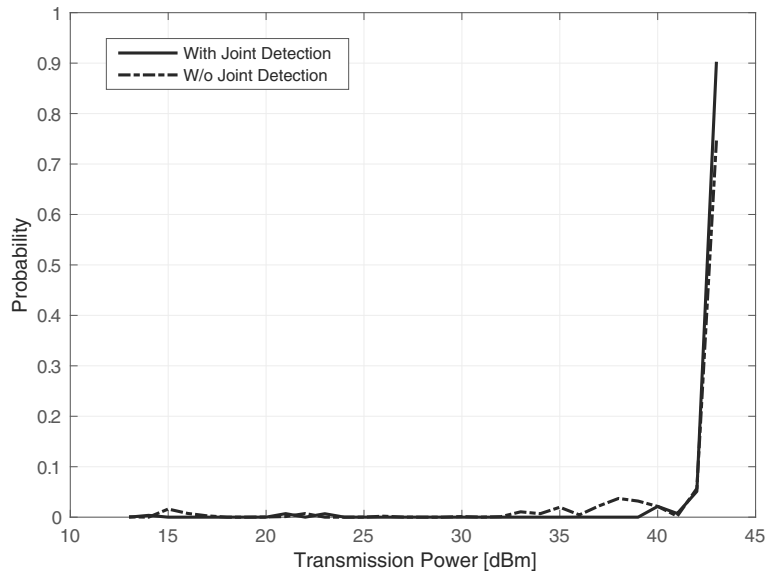
Fig. 13 Probability distribution of MBS transmission power (20 users/sector, PBS: 2(cell radius)/3)

Appendix A: Derivation of throughput
A.1 Throughput with joint ML detection
In this study, the constellation constraint capacity (CCC) is calculated as the throughput of the system. Suppose that the received signal on the lth subcarrier is given as:

$$y^l = |h^l_d| \exp\left(j\theta^l_d\right) x^l_d(k_d) + \left|h^l_i\right| x^l_i(k_i) + z^l, \quad (17)$$

where y^l is the received signal, $\left|h^l_d\right| \exp\left(j\theta^l_d\right)$ is the channel response between a receiver and a base station that transmits a desired signal, $|h^l_i|$ is the channel response between the receiver and a base station that causes interference, $\exp(j\theta^l_d)$ is the relative phase difference between h^l_d and h^l_i, $x^l_d(k_d)$ is the symbol with the k_dth constellation point of the desired signal, $x^l_i(k_i)$ is the symbol with the k_ith constellation point of the interference signal, and z^l is the AWGN. The CCC for the joint ML detection of $x_d(k_d)$ and $x_i(k_i)$ is calculated in the same way as in [22]. When $x^l_d(k_d)$ and $x^l_i(k_i)$ are received, the distribution of the received signal is equal to that of the AWGN, z^l, and it is given as:

$$p(z^l) = p\left(y^l \mid h^l_d| \exp\left(j\theta^l_d\right) x^l_d(k_d), \left|h^l_i\right| x^l_i\right)$$
$$= \exp\left[-|y^l - \left(\left|h^l_d\right|\exp\left(j\theta^l_d\right) x^l_d(k_d) + \left|h^l_i\right| x^l_i(k_i)|^2/2\sigma^2\right)\right] \cdot \left(2\pi\sigma^2\right)^{-1}$$
$$= \exp\left[-\left|z^l\right|^2/2\sigma^2\right] \cdot \left(2\pi\sigma^2\right)^{-1}, \quad (18)$$

where σ^2 is the variance of the AWGN. Joint ML detection calculates a likelihood with the knowledge of channel responses and modulation orders for both the desired and interference signals as follows. The likelihood for the k_dth constellation point of the desired signal is given by:

$$p\left(y^l \mid \left|h^l_d\right| x^l_d(k_d)\right) = \frac{1}{N_i} \sum_{i_i=0}^{N_i-1} p\left(y^l \mid \left|h^l_d\right| \exp\left(j\theta^l_d\right) x^l_d(k_d), \left|h^l_i\right| x^l_i(k_i)\right)$$
$$= \frac{1}{N_i} \sum_{i_i=0}^{N_i-1} \exp\left[-|y^l - \left(\left|h^l_d\right| \exp\left(j\theta^l_d\right) x^l_d(k_d)\right.\right.$$
$$\left.\left.+ \left|h^l_i\right| x^l_i(k_i)|^2/2\sigma^2\right] \cdot \left(2\pi\sigma^2\right)^{-1},$$
$$(19)$$

where N_i is the number of constellation points in the interference signal. The probability density function of y^l is then given by:

$$p(y^l) = \frac{1}{N_d N_i} \sum_{i_d=0}^{N_d-1} \sum_{i_i=0}^{N_i-1} p\left(y^l \mid \left|h^l_d\right| \exp\left(j\theta^l_d\right) x^l_d(k_d), \left|h^l_i\right| x^l_i(k_i)\right)$$
$$= \frac{1}{N_d N_i} \sum_{i_d=0}^{N_d-1} \sum_{i_i=0}^{N_i-1} \exp\left[-\left|y^l - \left(|h^l_d| \exp\left(j\theta^l_d\right) x^l_d(k_d)\right.\right.\right.$$
$$\left.\left.\left.+ \left|h^l_i\right| x^l_i(k_i)|^2/2\sigma^2\right.\right] \cdot \left(2\pi\sigma^2\right)^{-1}$$
$$= \frac{1}{N_d N_i} \sum_{i_d=0}^{N_d-1} \sum_{i_i=0}^{N_i-1} \exp\left[-\left|\left\{\left|h^l_d\right| \exp\left(j\theta^l_d\right) x^l_d(k_d) + \left|h^l_i\right| x^l_i(k_i)\right\}\right.\right.$$
$$\left.\left.-\left\{\left|h^l_d\right| \exp\left(j\theta^l_d\right) x^l_d(k_d) + \left|h^l_i\right| x^l_i(i_i)\right\} + z|^2/2\sigma^2\right] \cdot \left(2\pi\sigma^2\right)^{-1},$$
$$(20)$$

where N_d is the number of constellation points in the desired signal.

Using Eqs. (18) and (20), the CCC is calculated as:

$$
\left\{ C_{Wt}\left(y^l \mid h_d^l, h_i^l, N_d, N_i\right) \right\}
$$

$$
= \sum_{k_d=0}^{N_d-1}\sum_{k_i=0}^{N_i-1} \frac{1}{N_dN_i} \int_{\theta_d^l}\int_{-\infty}^{\infty} p\left(y^l \mid \left|h_d^l\right| \exp\left(j\theta_d^l\right) x_d^l(k_d),\ \left|h_i^l\right| x_i^l(k_i)\right)
$$

$$
\cdot \log_2\left(\frac{p\left(y^l \mid \left|h_d^l\right| \exp\left(j\theta_d^l\right) x_d^l(k_d),\ \left|h_i^l\right| x_i^l(k_i)\right)}{\frac{1}{N_dN_i}\sum_{i_d=0}^{N_d-1}\sum_{i_i=0}^{N_i-1} p\left(y^l \mid \left|h_d^l\right| \exp\left(j\theta_d^l\right) x_d^l(i_d),\ \left|h_i^l\right| x_i^l(i_i)\right)} \right) dy^l d\theta_d^l.
\tag{21}
$$

From [25], Eq. (21) is evaluated through Monte Carlo simulation, and the integral of y is given as:

$$
C_{Wt}\left(y^l \mid h_d^l, h_i^l, N_d, N_i\right)
$$

$$
= \log_2(N_dN_i)
$$

$$
- \frac{1}{N_dN_i}\sum_{k_d=0}^{N_n-1}\sum_{k_i=0}^{N_f-1}\mathop{E}_{\theta_d^l,z}
$$

$$
\left[\log_2\left(\frac{\sum_{i_d=0}^{N_d-1}\sum_{i_i=0}^{N_i-1} p\left(y^l \mid \left|h_d^l\right| \exp\left(j\theta_d^l\right) x_d^l(i_d),\ \left|h_i^l\right| x_i^l(i_i)\right)}{p\left(y^l \mid \left|h_d^l\right| \exp\left(j\theta_d^l\right) x_d^l(k_d),\ \left|h_i^l\right| x_i^l(i_i)\right)} \right) \right].
\tag{22}
$$

The first term of Eq. (22) is the amount of entropy for the $N_d \cdot N_i$ constellation points, and the second term is the average ambiguity of the received signal caused by the noise when joint ML detection is employed.

Using Eqs. (18) and (20), Eq. (22) is rewritten as:

$$
C_{Wt}\left(y^l \mid h_d^l, h_i^l, N_d, N_i\right) = \log_2(N_dN_i) - \frac{1}{N_dN_i}
$$

$$
\times \sum_{k_d=0}^{N_d-1}\sum_{k_i=0}^{N_i-1}\mathop{E}_{\theta_d^l,z}\left[\log_2\left(\sum_{i_d=0}^{N_d-1}\sum_{i_i=0}^{N_i-1} \exp\left(-\left|\left|h_d^l\right|\exp\left(j\theta_d^l\right)\left(x_d^l(k_d)-x_d^l(i_d)\right)\right.\right.\right.\right.
$$

$$
\left.\left.\left.\left. +\left|h_i^l\right|\left(\left(x_i^l(k_i)-x_i^l(i_i)\right)+z^l\right|^2/\sigma^2\right)\right) - \log_2\exp\left(-\left|z^l\right|^2/\sigma^2\right)\right].
\tag{23}
$$

The throughput obtained through joint ML detection can be given by the difference between the CCCs with and without the desired signal. This is calculated as:

$$
C_{Wt}\left(y^l \mid h_d^l, h_i^l, N_d, N_i\right) - C_{Wt}\left(\hat{y}^l \mid h_i^l, N_i\right)
$$

$$
= \log_2(N_dN_i) - \log_2(N_i) + \frac{1}{N_dN_i}\sum_{k_d=0}^{N_d-1}\sum_{k_i=0}^{N_i-1}\mathop{E}_{\theta_d^l,z}
$$

$$
\left[\log_2\left(\sum_{i_d=0}^{N_d-1}\sum_{i_i=0}^{N_i-1} \exp\left(-\left|\left|h_d^l\right|\exp\left(j\theta_d^l\right)\left(x_d^l(k_d)-x_d^l(i_d)\right)\right.\right.\right.\right.
$$

$$
\left.\left.\left. +\left|h_i^l\right|\left(\left(x_i^l(k_i)-x_i^l(i_i)\right)+z|^2/\sigma^2\right)\right)\right.
$$

$$
\left. -\log_2\left(\sum_{i_i=0}^{N_i-1}\exp\left(-\left|h_i^l\right|\left(x_i^l(k_i)-x_i^l(i_i)\right)+z\right|^2/\sigma^2\right) \right) \right],
\tag{24}
$$

where

$$
\hat{y}^l = |h_i^l| x_i^l(k_i) + z^l.
\tag{25}
$$

Equation (24) becomes Eq. (11) by setting $N_d = N_m$, $N_i = N_p$, $k_d = k_m$, $k_i = k_p$, $i_d = i_m$, $i_i = i_p$, $|h_d^l| = \alpha_M^l|h_{\mathrm{MU}_m}^l|$, $|h_i^l| = |h_{\mathrm{PU}_m}^l|$, $\theta_d^l = \theta_{U_m}^l$, $z^l = z_m^l$, $x_d^l(k_d^l) = x_{\mathrm{MU}_m}^l(k_m)$, and $x_i^l(k_i^l) = x_{\mathrm{PU}_p}^l(k_p)$. Similarly, by setting $N_d = N_p$, $N_i = N_m$, $k_d = k_p$, $k_i = k_m$, $i_d = i_p$, $i_i = i_m$, $|h_{\mathrm{PU}_p}^l|\exp\left(j\theta_{U_p}^l\right)$, $\left|h_i^l\right| = \alpha_M^l\left|h_{\mathrm{MU}_p}^l\right|$, $\theta_d^l = \theta_{U_p}^l$, $z^l = z_p^l$, $x_d^l(k_d^l) = x_{\mathrm{PU}_p}^l(k_p)$, and $x_i^l\left(k_i^l\right) = x_{\mathrm{MU}_m}^l(k_m)$, Eq. (24) becomes Eq. (12).

Throughput without joint ML detection

Equation (22) represents the CCC when both the desired and interference signals are detected. Without joint ML detection, only the desired signal is detected. The knowledge of the channel response and the modulation order for the desired signal is required, and the likelihood for the k_dth constellation point is given as follows:

$$
p(y^l \mid |h_d^l| x_d^l(k_d))
$$

$$
= p\left(y^l \mid \left|h_d^l\right| x_d^l(k_d),\ \left|h_i^l\right|\exp\left(j\theta_i^l\right) x_i^l(k_i)\right)
$$

$$
= \exp\left[-|y^l - \left(\left|h_d^l\right| x_d^l(k_d) + \left|h_i^l\right|\exp\left(j\theta_i^l\right) x_i^l(k_i)\right|^2/2\sigma^2 \right] \cdot \left(2\pi\sigma^2\right)^{-1}.
\tag{26}
$$

In this case, the CCC is given as:

$$
C_{Wo}\left(y^l \mid h_d^l, h_i^l, N_d, N_i\right) = \log_2(N_d) - \frac{1}{N_dN_i}
$$

$$
\times \sum_{k_d=0}^{N_d-1}\sum_{k_i=0}^{N_i-1}\mathop{E}_{\theta_d^l,z}\left[\log_2\left(\sum_{i_d=0}^{N_d-1}\exp\left(-\left|\left|h_d^l\right|\exp\left(j\theta_d^l\right)\left(x_d^l(k_d)-x_d^l(i_d)\right)\right.\right.\right.\right.
$$

$$
\left.\left. +\ \left|h_i^l\right| x_i^l(k_i) + z^l\right|^2/\sigma^2\right) - \log_2\exp\left(-\left|z^l\right|^2/\sigma^2\right) \right]
\tag{27}
$$

The throughput is also given as the tdifference between the CCCs with and without the desired signal. This is calculated in the same manner as:

$$
C_{Wo}\left(y^l \mid h_d^l, h_i^l, N_d, N_i\right) - C_{Wo}\left(\hat{y}^l \mid h_i^l, N_i\right) = \log_2(N_d) - \frac{1}{N_dN_i}
$$

$$
\times \sum_{k_d=0}^{N_d-1}\sum_{k_i=0}^{N_i-1}\mathop{E}_{\theta_d^l,z}\left[\log_2\left(\sum_{i_d=0}^{N_d-1}\exp\left(-\left|\left|h_d^l\right|\left(x_d^l(k_d)-x_d^l(i_d)\right)\right.\right.\right.\right.
$$

$$
\left.\left. +\ \left|h_i^l\right| x_i^l(k_i) + z^l\right|^2/\sigma^2\right) \right)
$$

$$
-\log_2\left(\sum_{i_i=0}^{N_i-1}\exp\left(-\left|\left|h_i^l\right| x_i^l(k_i) + z^l\right|^2/\sigma^2\right) \right) \right].
\tag{28}
$$

In the same way as in Sec. (5), by appropriate substitutions, Eq. (28) becomes Eqs. (9) and (10).

Abbreviations

AWGN: Additive white Gaussian noise; HetNet: Heterogeneous network; LTE: Long Term Evolution; MBS: Macro base station; MU: Macro user; MIMO: Multiple-input multiple-output; ML: Maximum likelihood; NOMA: Non-orthogonal multiple access; QAM: Quadrature amplitude shift keying; QPSK: Quadrature phase shift keying; PBS: Pico base station; PU: Pico user; PF:

Proportional fairness; RMS: Root mean square; SIC: Successive interference cancelation; SINR: Signal-to-interference-plus-noise ratio

Availability of data and materials
The first author have the data and source codes.

Authors' contributions
The author contributed to the proposal and evaluation of the written schemes. The author read and approved the final manuscript.

Authors' information
Yukitoshi Sanada received his B.E. degree in Electrical Engineering from the Keio University, Yokohama, Japan; his M.A.Sc. degree in Electrical Engineering from the University of Victoria, B.C., Canada; and his Ph.D. degree in Electrical Engineering from the Keio University, Yokohama, Japan, in 1992, 1995, and 1997, respectively. In 1997, he joined the Faculty of Engineering, Tokyo Institute of Technology, as a research associate. In 2000, he joined Advanced Telecommunication Laboratory, Sony Computer Science Laboratories, Inc., as an associate researcher. In 2001, he joined the Faculty of Science and Engineering, Keio University, where he is now a professor.

Competing interests
The author declares that he/she has no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References
1. International Telecommunication Union Radio (ITU-R), Assessment of the global mobile broadband deployments and forecasts for International Mobile Telecommunications. Report ITU-R M.2243 (2011)
2. H. Osada, M. Inamori, Y. Sanada, Non-orthogonal access scheme over multiple channels with iterative interference cancellation and fractional sampling in OFDM receiver. IEICE Trans. on Commun. **E95-B**(12), 3837–3844 (2012)
3. A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, T. Nakamura, Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access. Int. Symp. Intell. Signal Proc. Commun. Syst, 770–774 (2013)
4. Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, K. Higuchi, *Non-orthogonal multiple access (NOMA) for cellular future radio access. IEEE 77th Vehicular Technology Conference.*, (2013)
5. M.-R. Honjeij, J. Farah, C. A. Nour, C. Douillard, Resource allocation in downlink non-orthogonal multiple access NOMA for future radio access. IEEE 81st Vehicular Technology Conference (2015)
6. L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, Z. Wang, Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. IEEE Commun. Mag. **53**(9), 74–81 (2015)
7. Y. Saito, A. Benjebbour, Y. Kishiyama, T. Nakamura, System-level performance evaluation of downlink non-orthogonal multiple access (NOMA). IEEE 24th Int. Symp. Per. Indoor Mob. Radio Commun, 611–615 (2013)
8. K. Ando, Y. Sanada, Joint maximum likelihood detection in far user of non-orthogonal multiple access. IEEE VTS Asia Pacific Wireless Communications Symposium 2015 (2015)
9. K. Higuchi, A. Benjebbour, Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access. IEICE Trans. Commun. **E98-B**(3), 403–414 (2015)
10. M. B. Çelebi, H. Arslan, Theoretical analysis of the co-existence of LTE-A signals and design of an ML-SIC receiver. IEEE Trans. Wirel. Commun. **14**(8), 4626–4639 (2015)
11. E. Pateromichelakis, M. Shariat, A. ul Quddus, T. Tafazolli, On the evolution of multi-cell scheduling in 3GPP LTE / LTE-A. IEEE Commun. Surv. Tutor. **15**(2), 701–717 (2012)
12. Y. Peng, F. Qin, Exploring Het-Net in LTE-Advanced system: interference mitigation and performance improvement in macro-pico scenario. IEEE International Conference on Communications Workshops (2011)
13. A. Weber, O. Stanze, Scheduling strategies for HetNets using eICIC. IEEE Int. Conf. Commun, 6787–6791 (2012)
14. A. Weber, O. Stanze, Heterogeneous networks with LTE-Advanced technologies. Bell Labs Tech. J. **18**(1), 41–58 (2013)
15. A. Mokdad, P. Azmi, N. Mokari, Radio resource allocation for heterogeneous traffic in GFDM-NOMA heterogeneous cellular networks. IET Commun. **10**(12), 1444–1455 (2016)
16. P. Swami, V. Bhatia, S. Vuppala, T. Ratnarajah, User fairness and performance enhancement for cell edge user in NOMA-HCN with offloading. IEEE 85th Vehicular Technology Conference (2017)
17. M. Moltafet, N. Mokari, M. R. Javan, H. Saeedi, H. Pishro-Nik, A. Weber, O. Stanze, A new multiple access technique for 5G: power domain sparse code multiple access (PSMA). IEEE Access. **6**, 747–759 (2017)
18. T. Yazaki, Y. Sanada, Effect of joint detection and decoding in non-orthogonal multiple access. Wirel. Pers. Commun. **89**(4), 1203–1219 (2016)
19. T. Yazaki, Y. Sanada, Throughput performance of non-orthogonal multiple access with joint detection in far user. IEICE Trans. on Commun. **E100-B**(2), 344–353 (2017)
20. G. Ungerboeck, Channel coding with multilevel/phase signals. IEEE Trans. Info. Theory. **1**, 55–67 (1982)
21. J. Harshan, B. S. Rajan, A novel power allocation scheme for two-user GMAC with finite input constellations. IEEE Trans. Wirel. Commun. **12**(2), 818–827 (2013)
22. M. J. Hagh, M. Reza Soleymani, Constellation rotation for DVB multiple access channels with raptor coding. IEEE Trans. Broadcast. **59**(2), 290–297 (2013)
23. T. Seyama, T. Dateki, H. Seki, Efficient selection of user sets for downlink non-orthogonal multiple access. IEEE 26th Int. Symp. Per. Indoor Mob. Radio Commun, 1062–1066 (2015)
24. E. Mellios, Z. Mansor, G. S. Hilton, A. R. Nix, J. P. McGeehan, Impact of antenna pattern and handset rotation on macro-cell and pico-cell propagation in heterogeneous LTE networks. 2012 IEEE International Symposium on Antennas and Propagation (2012)
25. G. Ungerboeck, Channel coding with multilevel/phase signals. IEEE Trans. Info. Theory. **IT-28**(1), 55–67 (1982)