Short-Packet Communications for MIMO NOMA Systems over Nakagami-m Fading: BLER and Minimum Blocklength Analysis

Duc-Dung Tran, *Student Member, IEEE,* Shree Krishna Sharma, *Senior Member, IEEE,* Symeon Chatzinotas, *Senior Member, IEEE,* Isaac Woungang, *Senior Member, IEEE,* and Björn Ottersten, *Fellow, IEEE*

Abstract—Recently, ultra-reliable and low-latency communications (URLLC) using short-packets has been proposed to fulfill the stringent requirements regarding reliability and latency of emerging applications in 5G and beyond networks. In addition, multiple-input multiple-output non-orthogonal multiple access (MIMO NOMA) is a potential candidate to improve the spectral efficiency, reliability, latency, and connectivity of wireless systems. In this paper, we investigate short-packet communications (SPC) in a multiuser downlink MIMO NOMA system over Nakagami-m fading, and propose two antenna-user selection methods considering two clusters of users having different priority levels. In contrast to the widely-used long data-packet assumption, the SPC analysis requires the redesign of the communication protocols and novel performance metrics. Given this context, we analyze the SPC performance of MIMO NOMA systems using the average block error rate (BLER) and minimum blocklength, instead of the conventional metrics such as ergodic capacity and outage capacity. More specifically, to characterize the system performance regarding SPC, asymptotic (in the high signal-to-noise ratio regime) and approximate closed-form expressions of the average BLER at the users are derived. Based on the asymptotic behavior of the average BLER, an analysis of the diversity order, minimum blocklength, and optimal power allocation is carried out. The achieved results show that MIMO NOMA can serve multiple users simultaneously using a smaller blocklength compared with MIMO OMA, thus demonstrating the benefits of MIMO NOMA for SPC in minimizing the transmission latency. Furthermore, our results indicate that the proposed methods not only improve the BLER performance, but also guarantee full diversity gains for the respective users.

Index Terms—Block error rate, MIMO, minimum blocklength, non-orthogonal multiple access, short-packet communications.

I. INTRODUCTION

Ultra-reliable and low-latency communications (URLLC) has recently been considered as a promising technology for the 5th generation (5G) and beyond wireless networks to support novel applications with unprecedented requirements of reliability and latency [1–3]. Furthermore, it is a potential solution for mission-critical Internet of Things (IoT) applications such as industrial automation, remote surgery, and vehicle-to-everything (V2X) communications, which require

high reliability and low latency [4, 5]. URLLC systems should be designed to meet the requirements of high reliability (99.999%) and low latency (1 ms) [6]. To achieve such stringent requirements, a new transmission approach, i.e., short-packet communications (SPC), could be a promising solution. This is different from the traditional analytic methods designed to target Shannon's channel capacity using long data-packets, which are no longer suitable for low latency systems [6]. To characterize the performance of SPC, new performance metrics including block error rate (BLER) and overhead ratio (i.e., ratio of pilots to the information payload), have been introduced in the literature [7–9].

Besides, non-orthogonal multiple access (NOMA) has recently emerged as a promising technology to improve the spectral efficiency and user fairness for wireless networks [10, 11]. In contrast to the orthogonal multiple access (OMA) which utilizes orthogonal resources (e.g., time and frequency) to support multiple users, this technique can serve them at the same time/frequency/code by using the power domain and effective interference management methods, such as successive interference cancellation (SIC) [10]. Therefore, NOMA can more effectively support massive connectivity and further improve the reliability and latency for wireless systems [12, 13]. With its potential advantages, NOMA standardization has been recently studied in the 3rd Generation Partnership Project (3GPP) frameworks [14–16] including the 3GPP Release 16 [16]. Also, the latest trend is to employ NOMA in the uplink due to the emergence of IoT and machine-type communication systems [3, 16, 17]. Thus, NOMA and its variations are expected to be employed in various 5G and beyond application scenarios [13, 18, 19].

In addition, the combination of NOMA and multiple-input multiple-output (MIMO) systems (so-called MIMO NOMA), which can significantly enhance the spectral efficiency and performance of NOMA systems, has also been investigated in recent years [20, 21]. The ergodic capacity analysis of MIMO NOMA systems has been considered in [22], where the authors have proved the superiority of MIMO NOMA over MIMO OMA in terms of capacity. To exploit the spatial degrees of freedom, some MIMO NOMA schemes have been proposed in the literature [23, 24]. Specifically, the authors in [23] have considered a multi-beam MIMO NOMA scenario, where multiple analog beams are formed for downlink transmission of a NOMA user group by exploiting the channel sparsity and a large scale antenna array. Meanwhile, the work in [24]

has investigated a space-time coded MIMO NOMA system, where two users' signals are mapped into n-dimensional constellations corresponding to the same algebraic lattices from a number field. Although the system performance can be significantly improved with the increase in the utilized number of antennas in MIMO systems, this requires large power consumption and high complexity of signal processing [25]. To mitigate these issues while ensuring the diversity and capacity benefits from MIMO, transmit antenna selection (TAS) scheme has been proposed as a promising solution to improve the performance gain of MIMO NOMA systems [25–27]. It is noteworthy that the above works on MIMO NOMA have been conducted under the assumption of long data-packet transmissions, which is no longer applicable for emerging URLLC applications with short data-packets in 5G and beyond networks [3, 6, 28].

To overcome this challenge, in this paper, we propose to utilize SPC for MIMO NOMA systems to improve the reliability and latency as well as enhance the spectral efficiency and connectivity for wireless systems. As stated earlier, the large power consumption and high computational complexity of MIMO systems are putting a crucial challenge in designing effective communication protocols for SPC-based MIMO NOMA systems. Therefore, we consider a scenario, where TAS is used at the transmitter, and selection combining (SC) and maximal ratio combining (MRC) are utilized at the users with the purpose of improving the performance and reducing the complexity for MIMO NOMA systems with SPC. Herein, suitable performance metrics for SPC including average BLER and minimum blocklength, are utilized instead of the conventional ones such as ergodic capacity and outage capacity.

A. Related Works

Recently, there have been a few works on SPC in NOMA systems, which is considered as a promising solution to enhance the reliability, latency, and connectivity for wireless networks [29–36]. In particular, in [29], a two-user NOMA system with short-packets over Rayleigh fading channels was considered, in which the average BLER at users is derived to evaluate the system performance. In [30], the BLER performance of a NOMA system was addressed, where stochastic geometry and Nakagami-m fading channels are considered. In [31], X. Lai *et al.* analyzed the performance of a cooperative NOMA SPC system over Rayleigh fading channels. Furthermore, the transmission energy minimization problem and packet scheduling for two-user downlink NOMA systems with strictly heterogeneous latency constraints were investigated in [32, 33]. However, the works [29–33] only considered single-input single-output (SISO) systems.

To exploit the benefits of multiple antennas in improving the reliability and reducing the latency for SPC in NOMA systems, the work in [34] investigated a multiple-input single-output (MISO) scheme to evaluate the outage performance of a URLLC NOMA system with wireless power transfer. In [35], MIMO NOMA for URLLC systems was considered to enhance the reliability and latency performance of the

system. In this regard, a closed-form upper bound for the delay target violation probability was derived in [35] to identify the sufficient and necessary condition for the optimal transmit power. However, the analysis of average BLER and minimum blocklength was not considered in [35]. The works in both [34] and [35] investigated a scenario where an N-antenna base station (BS) provides services to N pairs of NOMA users, in which each pair of users is served by a distinct single transmit antenna. In contrast to this scenario, in [36], the combination of transmit antennas to serve a pair of users was examined in order to enhance the BLER performance of short-packet NOMA systems by utilizing the maximum ratio transmission (MRT), in which only the MISO scenario was considered.

Although MRT can significantly improve the system performance by combining all transmit antennas for transmission, it leads to high complexity of the signal processing and feedback overhead [37]. Against this context, TAS has been proposed as a low-complexity and power-efficient solution for multi-antenna transmitters to enhance the performance of NOMA systems by selecting a best transmit antenna for transmission that maximizes the signal-to-noise ratio (SNR) at the receiver side [25–27]. Nevertheless, the short-packet transmission in MIMO NOMA systems considering the TAS solution, average BLER, and minimum blocklength has not yet been analyzed. Furthermore, it is noted that most of these existing studies [29–31, 34–36] only investigated Rayleigh fading channels. Research on SPC for MIMO NOMA systems applying TAS for the transmitter, selection combining (SC) and maximal ratio combining (MRC) for the receiver, over a generic fading channel, i.e., Nakagami-m, to improve the system performance more effectively and bring more general insights of the system behavior has not yet been conducted, and thus is the focus of this paper.

B. Contributions

In contrast to the existing related works, in this paper, we propose a new framework to analyze the system performance of utilizing SPC in a NOMA network, in which MIMO and Nakagami-m distribution are considered. Most existing works on NOMA are conducted under the assumption that NOMA is carried out based on the difference in users' channel conditions [10–12, 19–22, 26, 27, 29–36]. More precisely, in a two-user downlink NOMA system, a BS transmits information to the users by superimposing users' messages with different transmit power levels [10]. The user having worse channel quality is allocated with the higher power level compared with the user having a better channel condition. However, in practice, users may have similar channel conditions but require different quality of service (QoS) as discussed in [38–40]. For example, some users may need to be served faster with low targeted data-rate, i.e, incident alerts, while some users can be served with the best effort, i.e., downloading of multimedia files [39]. In such a heterogeneous scenario, NOMA scheme becomes advantageous as compared to the conventional OMA as it can concurrently serve users having different QoS priorities with the same resources (time/frequency/code).

Given this context, we examine a scenario, in which a BS communicates with two user clusters having different

priority levels over Nakagami-m fading channels, where the BS and all users are equipped with multiple antennas. Note that Nakagami-m with parameter m is described as a general distribution that can include the well-known Rayleigh distribution and approximate the Rician one with the parameter K, where $m = (K+1)^2/(2K+1)$ [41]. In contrast to [30], which considers Nakagami-m fading channels for the BLER derivation in SISO case, our analysis derives the BLER expression for a more general scenario, i.e., MIMO. Herein, different MIMO schemes are investigated to reduce the complexity of signal processing and exploit the benefits of multiple antennas in improving the system performance. Particularly, at the BS, TAS is utilized to select the best transmit antenna for transmission that maximizes the post-processed SNR at the receiver [37]. Besides, at the user-side, two different diversity techniques are investigated: 1) SC, which selects the best received signal branch for further processing; and 2) MRC, which combines all the received signal branches from receive antennas to maximize the output SNR.

In addition, as discussed in [23, 24, 26, 27], assigning all users in a system for the implementation of NOMA is difficult due to the strong co-channel interference, leading to large complexity and high decoding delay. To overcome this issue, hybrid NOMA has recently been considered as a promising solution for 5G and beyond networks [42, 43]. Particularly, in this solution, all users in a network are divided into multiple small groups. Herein, the users in each group are served by NOMA, whereas the different groups are assigned to different orthogonal resource blocks (e.g., time or frequency). Therefore, in this paper, we consider a scenario, where users are paired[1] to perform NOMA with the purpose of decreasing the strong co-channel interference in NOMA systems [44–46]. This is a common assumption widely adopted in the NOMA literature to reduce the computational complexity and time delay of SIC decoding [23, 24, 26, 27]. It is noted that the achieved results from this analysis can be straightforwardly applied to different groups, which are incorporated into the network in an orthogonal manner.

Therefore, the main contributions of this paper are summarized as follows:

- Firstly, we propose a novel framework to analyze the performance of an SPC-based NOMA system, where MIMO transmission and Nakagami-m fading are considered. To achieve a general insight into the system behavior, we investigate two different cases of applying MIMO schemes for the transmitter and receiver sides including TAS/SC and TAS/MRC. Moreover, we investigate two antenna-user selection methods, namely high-priority cluster selection (HCS) and low-priority cluster selection (LCS), to design the effective communication protocols for SPC in a MIMO NOMA system.
- Secondly, we derive closed-form expressions for the average BLER of users in all considered cases. It should be noted that this work analyzes the performance in

[1] It is noted that the proposed schemes can be applied to the general scenario with more than two users within a NOMA group, which, however, results in higher computational complexity and larger time delay of SIC decoding, and is thus left for future work.

TABLE I
MAIN NOTATIONS AND SYMBOLS

Notation	Description
$\lvert\cdot\rvert$ and $\lVert\cdot\rVert$	The absolute value and the Euclidean norm
$\mathcal{CN}(0, N_0)$	A scalar complex Gaussian distribution with zero mean and variance N_0
$\mathbb{E}[\cdot]$	The expectation operator
$Q(x)$	The Gaussian Q-function
$\mathrm{Ei}(x)$	The exponential integral function
$\Gamma(x,t)$	The lower incomplete Gamma function
γ_0	Average transmit signal-to-noise ratio (SNR)
K_S and K_A	Number of antennas at base station and user A
n_A	Number of information bits for user A
N_A	Blocklength for user A
α_A	Power allocation coefficient for user A
D_A	Diversity order at user A
$\bar{\varepsilon}_A$	Average BLER at user A

terms of average BLER, which is more suitable for SPC than widely-used performance metrics such as ergodic capacity and outage capacity [6, 7].
- Thirdly, we derive asymptotic expressions for the average BLER in the high SNR regime and carry out an analysis of diversity order, minimum blocklength and optimal power allocation for SPC-based MIMO NOMA system based on the asymptotic average BLER.
- Finally, we perform the blocklength comparison between MIMO NOMA and MIMO OMA systems to clarify the superiority of MIMO NOMA compared to MIMO OMA in terms of low-latency transmission when considering SPC.

C. Paper Structure and Notations

The remainder of the paper is organized as follows. Section II depicts the system model and the proposed schemes in detail. Section III presents the performance analysis in terms of average BLER for the investigated scenarios. Section IV describes an analysis of the asymptotic average BLER, diversity order, optimal power allocation, and minimum blocklength. Section V presents the numerical results. Finally, Section VI concludes this paper. For clarity, we provide a summary of main notations and symbols used in this paper in Table I.

II. SYSTEM MODEL

In this paper, the SPC in a multiuser downlink MIMO NOMA system over Nakagami-m fading channels is considered, as depicted in Fig. 1. The network consists of one base station (BS), denoted by S, two cluster of users, denoted by $H = \{H_1, \ldots, H_I\}$ and $L = \{L_1, \ldots, L_J\}$. In addition, the BS and the users in both clusters H and L are equipped with K_S, K_H, and K_L antennas, respectively. As reported earlier in Section I, it is assumed that the users' QoS requirements are taken into account in the design of the MIMO NOMA transmission in SPC instead of their channel conditions. More precisely, we consider the scenario where the users in clusters H and L are treated as high-priority and low-priority ones, respectively. Furthermore, the users are paired to perform NOMA with the purpose of decreasing the strong co-channel interference in NOMA systems [44–46]. Specifically, each
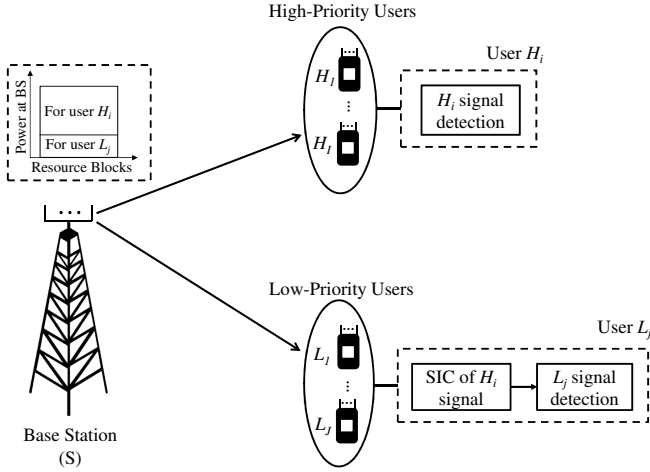
Fig. 1. Model of a MIMO NOMA system under SPC over Nakagami-m fading.

user pair consists of two users having different priorities selected from both the clusters H and L. Moreover, as mentioned earlier in Section I, to exploit the benefits of multiple antennas, we consider the scenario where TAS is employed at BS S whereas SC or MRC is utilized at the users' side (i.e., TAS/SC or TAS/MRC).

Regarding channel model, let $h_{S_k H_{i,r}}$ $\left(h_{S_k L_{j,s}}\right)$ $(1 \leq k \leq K_S, 1 \leq i \leq I, 1 \leq j \leq J, 1 \leq r \leq K_H, 1 \leq s \leq K_L)$ denote the channel coefficient of the link from antenna k at BS S to antenna r (s) at the user H_i (L_j). Herein, $h_{S_k H_{i,r}}$ $\left(h_{S_k L_{j,s}}\right)$ is an independent identically distributed (i.i.d) random variable following Nakagami-m distribution with parameter m_H (m_L) and mean value $\Omega_H = \mathbb{E}\left[\left|h_{S_k H_{i,r}}\right|^2\right]$ $\left(\Omega_L = \mathbb{E}\left[\left|h_{S_k L_{j,s}}\right|^2\right]\right)$. Thus, the Nakagami-$m$ distributions of $h_{S_k H_{i,r}}$ and $h_{S_k L_{j,s}}$ are, respectively, given by [41]

$$f_{h_{S_k H_{i,r}}}(x) = \frac{2m_H^{m_H} x^{2m_H-1}}{\Gamma(m_H)\Omega_H} e^{-\frac{m_H x^2}{\Omega_H}}, \tag{1}$$

and

$$f_{h_{S_k L_{j,s}}}(x) = \frac{2m_L^{m_L} x^{2m_L-1}}{\Gamma(m_L)\Omega_L} e^{-\frac{m_L x^2}{\Omega_L}}. \tag{2}$$

A. Antenna and User Selection

In this subsection, we present the proposed solutions of selecting antennas and users. As stated earlier, the user pairing is utilized for designing the MIMO NOMA. Specifically, the best user in each cluster is selected to perform NOMA based on the channel gains of the link from BS S to the users in order to improve the performance of NOMA implementation[2] [47]. Furthermore, we investigate two different antenna-user selection methods, i.e., HCS and LCS, which aim to improve the performance for the users selected from clusters H and L, respectively. It is noted that this selection process can

[2]In fact, some other sophisticated user pairing methods may further improve the performance of SPC-based MIMO NOMA systems. However, it is beyond the scope of this paper.

be carried out prior to information transmission through a suitable signaling and channel state information (CSI) estimation method [27]. In addition, as in [20, 26, 27], the perfect CSI scenario is considered and the required partial CSI, i.e., the instantaneous channel power gains, for each method is assumed to be available at the BS.

1) HCS Method: Since the users in cluster H has higher priority than those in cluster L, this method focuses on improving the performance of the selected user in cluster H. In particular, HCS method aims to jointly select a transmit antenna and a user in cluster H to maximize the channel power gain of the link from the BS S to the selected user.

For the TAS/SC scheme, the indices of selected transmit antenna, $\hat{k}$, user and receive antenna selected from cluster H, $\hat{i}$ and $\hat{r}_H$, are given by [37, 40]

$$\left(\hat{k}, \hat{i}, \hat{r}_H\right) = \arg \max_{1 \leq k \leq K_S, 1 \leq i \leq I, 1 \leq r \leq K_H} \left\{\left|h_{S_k H_{i,r}}\right|^2\right\}, \tag{3}$$

and the indices of user and receive antenna selected from cluster L, $\hat{j}$ and $\hat{r}_L$, are expressed as

$$\left(\hat{j}, \hat{r}_L\right) = \arg \max_{1 \leq j \leq J, 1 \leq r \leq K_L} \left\{\left|h_{S_{\hat{k}} L_{j,r}}\right|^2\right\}. \tag{4}$$

For TAS/MRC, $\hat{k}$, $\hat{i}$, and $\hat{j}$ are given by [37]

$$\left(\hat{k}, \hat{i}\right) = \arg \max_{1 \leq k \leq N_S, 1 \leq i \leq I} \left\{\left\|\mathbf{h}_{S_k H_i}\right\|^2\right\}, \tag{5}$$

and

$$\hat{j} = \arg \max_{1 \leq j \leq J} \left\{\left\|\mathbf{h}_{S_{\hat{k}} L_j}\right\|^2\right\}, \tag{6}$$

where $\mathbf{h}_{S_k H_i}$ $\left(\mathbf{h}_{S_k L_j}\right)$ represents the $K_H \times 1$ $(K_L \times 1)$ channel vector of the link from antenna k at BS S to user H_i (L_j).

2) LCS Method: To improve the performance of the selected user in cluster L which has a lower priority, an antenna at BS S and a user in cluster L are jointly chosen for transmission to provide the best channel power gain of the link from BS S to the selected user. Mathematically, $\hat{k}$, $\hat{i}$, $\hat{j}$, $\hat{r}_H$, and $\hat{r}_L$ in this method can be expressed as follows:

For TAS/SC:

$$\begin{cases} \left(\hat{k}, \hat{j}, \hat{r}_L\right) = \arg \max_{1 \leq k \leq K_S, 1 \leq j \leq J, 1 \leq r \leq K_L} \left\{\left|h_{S_k L_{j,r}}\right|^2\right\}, \\ \left(\hat{i}, \hat{r}_H\right) = \arg \max_{1 \leq i \leq I, 1 \leq r \leq K_H} \left\{\left|h_{S_{\hat{k}} H_{i,r}}\right|^2\right\}, \end{cases} \tag{7}$$

and for TAS/MRC:

$$\begin{cases} \left(\hat{k}, \hat{j}\right) = \arg \max_{1 \leq k \leq N_S, 1 \leq j \leq J} \left\{\left\|\mathbf{h}_{S_k L_j}\right\|^2\right\}, \\ \hat{i} = \arg \max_{1 \leq i \leq I} \left\{\left\|\mathbf{h}_{S_{\hat{k}} H_i}\right\|^2\right\}, \end{cases} \tag{8}$$

B. Information Transmission Process and Channel Statistics

With the NOMA protocol, BS S transmits the mixed message [44]

$$x = \sqrt{P_S \alpha_{H_{\hat{i}}}} x_{H_{\hat{i}}} + \sqrt{P_S \alpha_{L_{\hat{j}}}} x_{L_{\hat{j}}} \tag{9}$$

to users $H_{\hat{i}}$ and $L_{\hat{j}}$. Herein, P_S is the total transmit power, $\alpha_{H_{\hat{i}}}$ and $\alpha_{L_{\hat{j}}}$ $(\alpha_{H_{\hat{i}}} + \alpha_{L_{\hat{j}}} = 1)$ denote the power allocation

coefficients, as well as $x_{H_{\hat{i}}}$ and $x_{L_{\hat{j}}}$ represent the messages for users $H_{\hat{i}}$ and $L_{\hat{j}}$, respectively. It is noted that $\alpha_{H_{\hat{i}}} > \alpha_{L_{\hat{j}}} > 0$ due to higher priority of user $H_{\hat{i}}$. Thus, the received signal at user U $\left(U \in \left\{H_{\hat{i}}, L_{\hat{j}}\right\}\right)$ is given by

$$y_U = \mathbf{u}_U \mathbf{h}_{S_{\hat{k}}U} \sqrt{P_S} \left(\sqrt{\alpha_{H_{\hat{i}}}} x_{H_{\hat{i}}} + \sqrt{\alpha_{L_{\hat{j}}}} x_{L_{\hat{j}}}\right) + \mathbf{u}_U \mathbf{w}_U, \tag{10}$$

where $\mathbf{w}_U \sim \mathcal{CN}(0, N_0)$ denotes the additive white Gaussian noise (AWGN) at user U, and $\mathbf{u}_U$ represents the signal processing operation at user U, which is defined as in [48]

$$\mathbf{u}_U = \begin{cases} \mathbf{e}_{K_U, \hat{r}_U}, & \text{for TAS/SC} \\ \dfrac{\mathbf{h}_{S_{\hat{k}}U}^\dagger}{\left\|\mathbf{h}_{S_{\hat{k}}U}\right\|}, & \text{for TAS/MRC} \end{cases}, \tag{11}$$

where $\mathbf{e}_{K,i}$ is a $1 \times K$ vector whose the i-th element is equal to 1, and the others are zeros.

According to NOMA principle, the user $H_{\hat{i}}$ can directly decode its own message, $x_{H_{\hat{i}}}$, since it is allocated with larger transmit power (i.e., $\alpha_{H_{\hat{i}}} > \alpha_{L_{\hat{j}}}$), hence, the interference generated by the signal of the user $L_{\hat{j}}$, $x_{L_{\hat{j}}}$, can be treated as noise [27]. Thus, the instantaneous signal-to-interference-plus-noise ratio (SINR) at the user $H_{\hat{i}}$ to detect $x_{H_{\hat{i}}}$ is written as

$$\gamma_{H_{\hat{i}}}^{x_{H_{\hat{i}}}} = \frac{\alpha_{H_{\hat{i}}} \gamma_0 g_{SH}}{\alpha_{L_{\hat{j}}} \gamma_0 g_{SH} + 1}, \tag{12}$$

where $\gamma_0 = \frac{P_S}{N_0}$ denotes the average transmit SNR and g_{SH} is defined as

$$g_{SH} = \begin{cases} \left|h_{S_{\hat{k}}H_{\hat{i}, \hat{r}_H}}\right|^2, & \text{for TAS/SC} \\ \left\|\mathbf{h}_{S_{\hat{k}}H_{\hat{i}}}\right\|^2, & \text{for TAS/MRC} \end{cases}. \tag{13}$$

Meanwhile, given the considered NOMA scenario, where the user $H_{\hat{i}}$ is served with higher priority than the user $L_{\hat{j}}$ (i.e., $\alpha_{H_{\hat{i}}} > \alpha_{L_{\hat{j}}}$), the user $L_{\hat{j}}$ first needs to decode $x_{H_{\hat{i}}}$ and then remove this component from the received signal by using SIC before detecting its own message, $x_{L_{\hat{j}}}$, [27]. Unlike [27, 29–31] considering the perfect SIC (PSIC), in this paper, we consider the imperfect SIC (ISIC) scenario[3] to achieve more practical insights, where there exists a residual interference component due to the ISIC process [49]. Thus, the instantaneous SINRs at the user $L_{\hat{j}}$ to detect $x_{H_{\hat{i}}}$ and $x_{L_{\hat{j}}}$ are respectively expressed as

$$\gamma_{L_{\hat{j}}}^{x_{H_{\hat{i}}}} = \frac{\alpha_{H_{\hat{i}}} \gamma_0 g_{SL}}{\alpha_{L_{\hat{j}}} \gamma_0 g_{SL} + 1}, \tag{14}$$

and

$$\gamma_{L_{\hat{j}}}^{x_{L_{\hat{j}}}} = \frac{\alpha_{L_{\hat{j}}} \gamma_0 g_{SL}}{\psi \alpha_{H_{\hat{i}}} \gamma_0 g_{SL} + 1}, \tag{15}$$

where $\psi = \mathbb{E}\left[\left|x_{H_{\hat{i}}} - \hat{x}_{H_{\hat{i}}}\right|^2\right]$ denotes the level of residual interference caused by the ISIC process at user $L_{\hat{j}}$, which indicates the difference between the actual signal $x_{H_{\hat{i}}}$ and the estimated signal $\hat{x}_{H_{\hat{i}}}$. Specifically, $\psi = 0$ means perfect SIC

[3]It is noteworthy to mention that in this paper, we consider perfect CSI to evaluate the effects of ISIC on the performance of SPC-based MIMO NOMA systems. However, analyzing the impact of imperfect CSI on the performance of the SPC systems is also an important problem to be investigated in future works.

and $0 < \psi \le 1$ denotes ISIC. In (14) and (15), g_{SL} is given by

$$g_{SL} = \begin{cases} \left|h_{S_{\hat{k}}L_{\hat{j}, \hat{r}_L}}\right|^2, & \text{for TAS/SC} \\ \left\|\mathbf{h}_{S_{\hat{k}}L_{\hat{j}}}\right\|^2, & \text{for TAS/MRC} \end{cases}. \tag{16}$$

III. PROPOSED APPROACH FOR BLER PERFORMANCE ANALYSIS WITH SPC

In this section, we present some preliminaries on SPC and average BLER calculation, the derivation of CDF of channel power gains, and the average BLER analysis by utilizing HCS and LCS methods with TAS/SC and TAS/MRC schemes, specified in Section II-A.

A. Preliminaries

Considering SPC with blocklength N, the Shannon capacity C, and the BLER ε, the maximum achievable rate can be expressed as [29]:

$$R = C - \sqrt{\frac{v}{N}} Q^{-1}(\varepsilon) + O\left(\frac{\log_2 N}{N}\right), \tag{17}$$

where $Q^{-1}(x)$ is the inverse of the Gaussian Q-function, $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$, $C = \log_2(1 + \gamma)$, $v = (\log_2 e)^2 \left[1 - \frac{1}{(1+\gamma)^2}\right]$ represents the channel dispersion, γ is the SNR or SINR, and $O\left(\frac{\log_2 N}{N}\right)$ is the remainder terms of order $\frac{\log_2 N}{N}$. From (17), an approximation method, which is commonly referred to as normal approximation [7], is utilized to compute the instantaneous BLER as follows:

$$\varepsilon \approx Q\left(\frac{C - n/N}{\sqrt{v/N}}\right), \tag{18}$$

where $R = n/N$, n denotes the number of information bits, and the approximation is achieved by omitting the term $O\left(\frac{\log_2 N}{N}\right)$ when $N \ge 100$ as in [7].

Based on (18), the instantaneous BLER of decoding the message of user V, x_V $\left(V \in \left\{H_{\hat{i}}, L_{\hat{j}}\right\}\right)$, at user U $\left(U \in \left\{H_{\hat{i}}, L_{\hat{j}}\right\}\right)$ is given by [29]:

$$\varepsilon_U^{x_V} \approx Q\left(\frac{\log_2(1 + \gamma_U^{x_V}) - n_V/N_V}{\sqrt{v_U^{x_V}/N_V}}\right), \tag{19}$$

where $v_U^{x_V} = (\log_2 e)^2 \left[1 - \frac{1}{(1+\gamma_U^{x_V})^2}\right]$, n_V and N_V denote the number of information bits and blocklength to user V, respectively. Thus, the instantaneous BLER is calculated through the received SINR, the Shannon capacity, the number of information bits, and the blocklength; and specific encoding and modulation methods are not considered. From (19), the average BLER $\bar{\varepsilon}_U^{x_V}$ has the following form

$$\bar{\varepsilon}_U^{x_V} \approx \int_0^\infty \varepsilon_U^{x_V} f_{\gamma_U^{x_V}}(x) \, dx, \tag{20}$$

where $f_X(x)$ is the probability density function (PDF) of a random variable X. It is challenging to derive $\bar{\varepsilon}_U^{x_V}$ in (20). Therefore, an approximation[4] of $\varepsilon_U^{x_V}$ is utilized as discussed in [50], i.e.,

$$\varepsilon_U^{x_V} \approx \begin{cases} 1, & \gamma_U^{x_V} \leq v_V \\ A_U^{x_V}, & v_V < \gamma_U^{x_V} < \mu_V \\ 0, & \gamma_U^{x_V} \geq \mu_V \end{cases}, \quad (21)$$

where $A_U^{x_V} = 0.5 - \chi_V \sqrt{N_V} (\gamma_U^{x_V} - \beta_V)$, $\chi_V = \sqrt{\frac{1}{2\pi \left(2^{\frac{2n_V}{N_V}} - 1 \right)}}$, $v_V = \beta_V - \frac{1}{2\chi_V\sqrt{N_V}}$, $\mu_V = \beta_V + \frac{1}{2\chi_V\sqrt{N_V}}$, and $\beta_V = 2^{\frac{n_V}{N_V}} - 1$. By substituting (21) into (20), $\bar{\varepsilon}_U^{x_V}$ can be rewritten as

$$\bar{\varepsilon}_U^{x_V} \approx \chi_V \sqrt{N_V} \int_{v_V}^{\mu_V} F_{\gamma_U^{x_V}}(x)\, dx. \quad (22)$$

For user $H_{\hat{i}}$, from (12) and (22), its average BLER is expressed as

$$\begin{aligned} \bar{\varepsilon}_{H_{\hat{i}}} &= \bar{\varepsilon}_{H_{\hat{i}}}^{x_{H_{\hat{i}}}} \\ &\approx \chi_{H_{\hat{i}}} \sqrt{N_{H_{\hat{i}}}} \int_{v_{H_{\hat{i}}}}^{\mu_{H_{\hat{i}}}} F_{\gamma_{H_{\hat{i}}}^{x_{H_{\hat{i}}}}}(x)\, dx. \end{aligned} \quad (23)$$

For user $L_{\hat{j}}$, it first needs to remove the message of user $H_{\hat{i}}$, i.e., $x_{H_{\hat{i}}}$, by using ISIC before detecting its own message, i.e., $x_{L_{\hat{j}}}$. Therefore, user $L_{\hat{j}}$ cannot decode $x_{L_{\hat{j}}}$ if it decodes $x_{H_{\hat{i}}}$ unsuccessfully. This will affect its BLER performance. Given this context, the average BLER at user $L_{\hat{j}}$ is given by

$$\bar{\varepsilon}_{L_{\hat{j}}} = \bar{\varepsilon}_{L_{\hat{j}}}^{x_{H_{\hat{i}}}} + \left(1 - \bar{\varepsilon}_{L_{\hat{j}}}^{x_{H_{\hat{i}}}} \right) \bar{\varepsilon}_{L_{\hat{j}}}^{x_{L_{\hat{j}}}}, \quad (24)$$

where

$$\bar{\varepsilon}_{L_{\hat{j}}}^{x_{H_{\hat{i}}}} \approx \chi_{H_{\hat{i}}} \sqrt{N_{H_{\hat{i}}}} \int_{v_{H_{\hat{i}}}}^{\mu_{H_{\hat{i}}}} F_{\gamma_{L_{\hat{j}}}^{x_{H_{\hat{i}}}}}(x)\, dx,$$

and

$$\bar{\varepsilon}_{L_{\hat{j}}}^{x_{L_{\hat{j}}}} \approx \chi_{L_{\hat{j}}} \sqrt{N_{L_{\hat{j}}}} \int_{v_{L_{\hat{j}}}}^{\mu_{L_{\hat{j}}}} F_{\gamma_{L_{\hat{j}}}^{x_{L_{\hat{j}}}}}(x)\, dx.$$

B. Derivation for Cumulative Distribution Function (CDF) of Channel Power Gains

To derive the average BLER at users $H_{\hat{i}}$ and $L_{\hat{j}}$, we first need to calculate the CDFs of g_{SH} and g_{SL} with TAS/SC and TAS/MRC schemes in both HCS and LCS methods. Note that these derivations are based on the MIMO diversity techniques, channel distribution, and the antenna-user selection methods utilized for the analysis, regardless of the types of transmission (e.g., SPC or long data-packet transmissions). This is described as follows:

[4]In this paper, we consider an approximation method for BLER, as discussed in [50], to analyze the performance of SPC-based MIMO NOMA systems in terms of BLER. Deriving the error bound for BLER based on Jensen's inequality [51], which is more challenging, could be an interesting problem to investigate in future work.

1) HCS Method: The CDFs of g_{SH} and g_{SL} with HCS method are derived in the following propositions.

Proposition 1. *Under HCS method and Nakagami-m fading, the CDF of g_{SH} with TAS/SC and TAS/MRC schemes is given by*

$$F_{g_{SH}}^{HCS}(x) = 1 + \sum_{p=1}^{a_{H,I}} \sum_{\Delta_H = p} \Phi_H c_{H,I} x^{\varphi_H} e^{-\frac{p m_H x}{\lambda_{SH}}}, \quad (25)$$

where $\Delta_H = \sum_{q=0}^{b_H-1} \delta_{H,q}$, $\varphi_H = \sum_{q=0}^{b_H-1} q\delta_{H,q}$, $\Phi_H = (-1)^p \left[\prod_{q=0}^{b_H-1} \left(\frac{m_H^q}{q! \lambda_{SH}^q} \right)^{\delta_{H,q}} \right]$, $\lambda_{SH} = \Omega_H d_{SH}^{-\theta}$, $a_{H,I} = \begin{cases} K_S K_H I, & \text{for TAS/SC} \\ K_S I, & \text{for TAS/MRC} \end{cases}$, $b_H = \begin{cases} m_H, & \text{for TAS/SC} \\ m_H K_H, & \text{for TAS/MRC} \end{cases}$, *and* $c_{H,I} = \begin{pmatrix} a_{H,I} \\ p \end{pmatrix} \begin{pmatrix} p \\ \delta_{H,0}, \ldots, \delta_{H,b_H-1} \end{pmatrix}$, d_{SH} *and* θ *denote the distance and path loss exponent of the link from BS S to user $H_{\hat{i}}$, respectively.*

Proof: See Appendix A. ∎

Proposition 2. *Under HCS method and Nakagami-m fading, the CDF of g_{SL} with TAS/SC and TAS/MRC schemes is expressed as*

$$F_{g_{SL}}^{HCS}(x) = 1 + \sum_{p=1}^{a_{L,I}} \sum_{\Delta_L = p} \Phi_L c_{L,I} x^{\varphi_L} e^{-\frac{p m_L x}{\lambda_{SL}}}, \quad (26)$$

where $\Delta_L = \sum_{q=0}^{b_L-1} \delta_{L,q}$, $\varphi_L = \sum_{q=0}^{b_L-1} q\delta_{L,q}$, $\Phi_L = (-1)^p \left[\prod_{q=0}^{b_L-1} \left(\frac{m_L^q}{q! \lambda_{SL}^q} \right)^{\delta_{L,q}} \right]$, $\lambda_{SL} = \Omega_L d_{SL}^{-\theta}$, $a_{L,I} = \begin{cases} K_L J, & \text{for TAS/SC} \\ J, & \text{for TAS/MRC} \end{cases}$, $b_L = \begin{cases} m_L, & \text{for TAS/SC} \\ m_L K_L, & \text{for TAS/MRC} \end{cases}$, *and* $c_{L,I} = \begin{pmatrix} a_{L,I} \\ p \end{pmatrix} \begin{pmatrix} p \\ \delta_{L,0}, \ldots, \delta_{L,b_L-1} \end{pmatrix}$, d_{SL} *and* θ *denotes the distance of the link from BS S to user $L_{\hat{j}}$, respectively.*

Proof: It is noted that TAS is used to select the best transmit antenna for user $H_{\hat{i}}$ in this case, hence, it is considered as a random solution for user $L_{\hat{j}}$. As such, using (3), (4), (5), and (6), the CDF of g_{SL} is given by [37, 52]

$$F_{g_{SL}}(x) = \left(1 - \sum_{p=0}^{b_L-1} \frac{m_L^p}{p! \lambda_{SL}^p} x^p e^{-\frac{m_L x}{\lambda_{SL}}} \right)^{a_{L,I}}. \quad (27)$$

By using binomial expansion and multinomial theorem similar to the proof of Proposition 1 in Appendix A, we obtain the final expression of $F_{g_{SL}}(x)$ as in (26) and the proof is completed. ∎

2) LCS Method: Utilizing (7), (8), and algebraic manipulations similar to the proof of Proposition 1 in Appendix A, the CDF of g_{SH} and g_{SL} in this case are expressed as

$$F_{g_{SH}}^{LCS}(x) = 1 + \sum_{p=1}^{a_{H,II}} \sum_{\Delta_H = p} \Phi_H c_{H,II} x^{\varphi_H} e^{-\frac{p m_H x}{\lambda_{SH}}}, \quad (28)$$

and

$$F_{gSL}^{LCS}(x) = 1 + \sum_{p=1}^{a_{L,II}} \sum_{\Delta_L = p} \Phi_L c_{L,II} x^{\varphi_L} e^{-\frac{pm_L x}{\lambda_{SL}}}, \quad (29)$$

where
$$a_{H,II} = \begin{cases} K_H I, & \text{for TAS/SC} \\ I, & \text{for TAS/MRC} \end{cases},$$
$$c_{H,II} = \begin{pmatrix} a_{H,II} \\ p \end{pmatrix} \begin{pmatrix} p \\ \delta_{H,0}, \dots, \delta_{H,b_H-1} \end{pmatrix},$$
$$a_{L,II} = \begin{cases} K_S K_L J, & \text{for TAS/SC} \\ K_S J, & \text{for TAS/MRC} \end{cases}, \quad \text{and}$$
$$c_{L,II} = \begin{pmatrix} a_{L,II} \\ p \end{pmatrix} \begin{pmatrix} p \\ \delta_{L,0}, \dots, \delta_{L,b_L-1} \end{pmatrix}.$$

C. Average BLER Analysis of HCS Method

The derivation of the average BLER at users $H_{\hat{i}}$ and $L_{\hat{j}}$ in case of using the TAS/SC or TAS/MRC scheme with HCS method are provided in the following theorems.

Theorem 1. *Under HCS method and Nakagami-m fading, the average BLER at user $H_{\hat{i}}$ utilizing TAS/SC or TAS/MRC is expressed as*

$$\bar{\varepsilon}_{H_{\hat{i}}}^{HCS} \approx 1 + \frac{\chi_{H_{\hat{i}}} \alpha_{H_{\hat{i}}} \sqrt{N_{H_{\hat{i}}}}}{\gamma_0 \alpha_{L_{\hat{j}}}^2} \sum_{p=1}^{a_{H,I}} \sum_{\Delta_H = p} \sum_{q=0}^{\varphi_H} \begin{pmatrix} \varphi_H \\ q \end{pmatrix}$$
$$\times \left(-\frac{1}{\gamma_0 \alpha_{L_{\hat{j}}}} \right)^q \Phi_H c_{H,I} e^{\frac{\omega_H}{\gamma_0 \alpha_{L_{\hat{j}}}}} \mathcal{A}_H, \quad (30)$$

where

$$\mathcal{A}_H = \begin{cases} \omega_H \Xi_{H,1} + \Xi_{H,2}, & \hat{\varphi}_H = -2 \\ -\Xi_{H,1}, & \hat{\varphi}_H = -1 \\ \omega_H^{-\hat{\varphi}_H - 1} \Xi_{H,3}, & \hat{\varphi}_H \geq 0 \end{cases},$$

$\omega_H = \frac{pm_H}{\lambda_{SH}}$, $\Xi_{H,1} = \text{Ei}\left(-\omega_H \phi_{H_{\hat{i}}}\right) - \text{Ei}\left(-\omega_H \kappa_{H_{\hat{i}}}\right)$, $\Xi_{H,2} = \frac{e^{-\omega_H \phi_{H_{\hat{i}}}}}{\phi_{H_{\hat{i}}}} - \frac{e^{-\omega_H \kappa_{H_{\hat{i}}}}}{\kappa_{H_{\hat{i}}}}$, $\Xi_{H,3} = \Gamma\left(\hat{\varphi}_H + 1, \omega_H \phi_{H_{\hat{i}}}\right) - \Gamma\left(\hat{\varphi}_H + 1, \omega_H \kappa_{H_{\hat{i}}}\right)$, $\phi_{H_{\hat{i}}} = \frac{1}{\gamma_0 \alpha_{L_{\hat{j}}}} + B_{v_{H_{\hat{i}}}}$, $\kappa_{H_{\hat{i}}} = \frac{1}{\gamma_0 \alpha_{L_{\hat{j}}}} + B_{\mu_{H_{\hat{i}}}}$, $B_x = \frac{x}{\gamma_0 \left(\alpha_{H_{\hat{i}}} - \alpha_{L_{\hat{j}}} x\right)}$, and $\hat{\varphi}_H = \varphi_H - q - 2$.

Proof: See Appendix B. ∎

Theorem 2. *Under HCS method and Nakagami-m fading, the average BLER at user $L_{\hat{j}}$ utilizing TAS/SC or TAS/MRC is given by*

$$\bar{\varepsilon}_{L_{\hat{j}}}^{HCS} = \bar{\varepsilon}_{L_{\hat{j}}}^{x_{H_{\hat{i}}}, HCS} + \left(1 - \bar{\varepsilon}_{L_{\hat{j}}}^{x_{H_{\hat{i}}}, HCS}\right) \bar{\varepsilon}_{L_{\hat{j}}}^{x_{L_{\hat{j}}}, HCS}, \quad (31)$$

where

$$\bar{\varepsilon}_{L_{\hat{j}}}^{x_{H_{\hat{i}}}, HCS} \approx 1 + \frac{\chi_{H_{\hat{i}}} \alpha_{H_{\hat{i}}} \sqrt{N_{H_{\hat{i}}}}}{\gamma_0 \alpha_{L_{\hat{j}}}^2} \sum_{p=1}^{a_{L,I}} \sum_{\Delta_L = p} \sum_{q=0}^{\varphi_L} \begin{pmatrix} \varphi_L \\ q \end{pmatrix}$$
$$\times \left(-\frac{1}{\gamma_0 \alpha_{L_{\hat{j}}}} \right)^q \Phi_L c_{L,I} e^{\frac{\omega_L}{\gamma_0 \alpha_{L_{\hat{j}}}}} \mathcal{A}_L \left(\phi_{H_{\hat{i}}}, \kappa_{H_{\hat{i}}}\right),$$

$$\bar{\varepsilon}_{L_{\hat{j}}}^{x_{L_{\hat{j}}}, HCS} = \begin{cases} \bar{\varepsilon}_{L_{\hat{j}}, 1}^{x_{L_{\hat{j}}}, HCS}, & \psi = 0 \\ \bar{\varepsilon}_{L_{\hat{j}}, 2}^{x_{L_{\hat{j}}}, HCS}, & 0 < \psi \leq 1 \end{cases},$$

$$\bar{\varepsilon}_{L_{\hat{j}}, 1}^{x_{L_{\hat{j}}}, HCS} \approx 1 + \chi_{L_{\hat{j}}} \sqrt{N_{L_{\hat{j}}}} \sum_{p=1}^{a_{L,I}} \sum_{\Delta_L = p} \frac{\Phi_L c_{L,I} \hat{\omega}_L^{-\varphi_L - 1}}{\left(\alpha_{L_{\hat{j}}} \gamma_0\right)^{\varphi_L}} \Xi_{L,4},$$

$$\bar{\varepsilon}_{L_{\hat{j}}, 2}^{x_{L_{\hat{j}}}, HCS} \approx 1 + \frac{\chi_{L_{\hat{j}}} \alpha_{L_{\hat{j}}} \sqrt{N_{L_{\hat{j}}}}}{\gamma_0 \psi^2 \alpha_{H_{\hat{i}}}^2} \sum_{p=1}^{a_{L,I}} \sum_{\Delta_L = p} \sum_{q=0}^{\varphi_L} \begin{pmatrix} \varphi_L \\ q \end{pmatrix}$$
$$\times \left(-\frac{1}{\gamma_0 \psi \alpha_{H_{\hat{i}}}} \right)^q \Phi_L c_{L,I} e^{\frac{\omega_L}{\gamma_0 \eta \alpha_{H_{\hat{i}}}}} \mathcal{A}_L \left(\phi_{L_{\hat{j}}}, \kappa_{L_{\hat{j}}}\right),$$

$$\mathcal{A}_L (x,y) = \begin{cases} \omega_L \Xi_{L,1}^{(x,y)} + \Xi_{L,2}^{(x,y)}, & \hat{\varphi}_L = -2 \\ -\Xi_{L,1}^{(x,y)}, & \hat{\varphi}_L = -1 \\ \omega_L^{-\hat{\varphi}_L - 1} \Xi_{L,3}^{(x,y)}, & \hat{\varphi}_L \geq 0 \end{cases},$$

$\Xi_{L,1}^{(x,y)} = \text{Ei}\left(-\omega_L x\right) - \text{Ei}\left(-\omega_L y\right)$, $\Xi_{L,2}^{(x,y)} = \frac{e^{-\omega_L x}}{x} - \frac{e^{-\omega_L y}}{y}$, $\Xi_{L,3}^{(x,y)} = \Gamma\left(\hat{\varphi}_L + 1, \omega_L x\right) - \Gamma\left(\hat{\varphi}_L + 1, \omega_L y\right)$, $x \in \left\{\phi_{H_{\hat{i}}}, \hat{\phi}_{L_{\hat{j}}}\right\}$, $y \in \left\{\kappa_{H_{\hat{i}}}, \hat{\kappa}_{L_{\hat{j}}}\right\}$, $\phi_{L_{\hat{j}}} = \frac{1}{\gamma_0 \psi \alpha_{H_{\hat{i}}}} + \hat{B}_{v_{L_{\hat{j}}}}$, $\kappa_{L_{\hat{j}}} = \frac{1}{\gamma_0 \psi \alpha_{H_{\hat{i}}}} + \hat{B}_{\mu_{L_{\hat{j}}}}$, $\hat{B}_z = \frac{z}{\gamma_0 \left(\alpha_{L_{\hat{j}}} - \psi \alpha_{H_{\hat{i}}} z\right)}$, $\Xi_{L,4} = \Gamma\left(\varphi_L + 1, \hat{\omega}_L v_{L_{\hat{j}}}\right) - \Gamma\left(\varphi_L + 1, \hat{\omega}_L \mu_{L_{\hat{j}}}\right)$, $\omega_L = \frac{pm_L}{\lambda_{SL}}$, $\hat{\varphi}_L = \varphi_L - q - 2$, and $\hat{\omega}_L = \frac{pm_L}{\lambda_{SL} \alpha_{L_{\hat{j}}} \gamma_0}$.

Proof: See Appendix C. ∎

D. Average BLER Analysis of LCS Method

In this case, the average BLER at user $H_{\hat{i}}$ and $L_{\hat{j}}$ are derived through the following theorems.

Theorem 3. *Under LCS method and Nakagami-m fading, the average BLER at user $H_{\hat{i}}$ with TAS/SC or TAS/MRC is expressed as*

$$\bar{\varepsilon}_{H_{\hat{i}}}^{LCS} \approx 1 + \frac{\chi_{H_{\hat{i}}} \alpha_{H_{\hat{i}}} \sqrt{N_{H_{\hat{i}}}}}{\gamma_0 \alpha_{L_{\hat{j}}}^2} \sum_{p=1}^{a_{H,II}} \sum_{\Delta_H = p} \sum_{q=0}^{\varphi_H} \begin{pmatrix} \varphi_H \\ q \end{pmatrix}$$
$$\times \left(-\frac{1}{\gamma_0 \alpha_{L_{\hat{j}}}} \right)^q \Phi_H c_{H,II} e^{\frac{\omega_H}{\gamma_0 \alpha_{L_{\hat{j}}}}} \mathcal{A}_H. \quad (32)$$

Proof: To derive $\bar{\varepsilon}_{H_{\hat{i}}}^{LCS}$ in this theorem, the algebraic manipulations similar to the derivation of $\bar{\varepsilon}_{H_{\hat{i}}}^{HCS}$ in Appendix B can be utilized, where (28) is employed instead of (25). ∎

Theorem 4. *Under LCS method and Nakagami-m fading, the average BLER at user $L_{\hat{j}}$ with TAS/SC or TAS/MRC is given by*

$$\bar{\varepsilon}_{L_{\hat{j}}}^{LCS} = \bar{\varepsilon}_{L_{\hat{j}}}^{x_{H_{\hat{i}}}, LCS} + \left(1 - \bar{\varepsilon}_{L_{\hat{j}}}^{x_{H_{\hat{i}}}, LCS}\right) \bar{\varepsilon}_{L_{\hat{j}}}^{x_{L_{\hat{j}}}, LCS}, \quad (33)$$

where

$$\bar{\varepsilon}_{L_{\hat{j}}}^{x_{H_{\hat{i}}}, LCS} \approx 1 + \frac{\chi_{H_{\hat{i}}} \alpha_{H_{\hat{i}}} \sqrt{N_{H_{\hat{i}}}}}{\gamma_0 \alpha_{L_{\hat{j}}}^2} \sum_{p=1}^{a_{L,II}} \sum_{\Delta_L = p} \sum_{q=0}^{\varphi_L} \begin{pmatrix} \varphi_L \\ q \end{pmatrix}$$
$$\times \left(-\frac{1}{\gamma_0 \alpha_{L_{\hat{j}}}} \right)^q \Phi_L c_{L,II} e^{\frac{\omega_L}{\gamma_0 \alpha_{L_{\hat{j}}}}} \mathcal{A}_L \left(\phi_{H_{\hat{i}}}, \kappa_{H_{\hat{i}}}\right),$$

$$\bar{\varepsilon}_{L_{\hat{j}}}^{x_{L_{\hat{j}}}, LCS} = \begin{cases} \bar{\varepsilon}_{L_{\hat{j}}, 1}^{x_{L_{\hat{j}}}, LCS}, & \psi = 0 \\ \bar{\varepsilon}_{L_{\hat{j}}, 2}^{x_{L_{\hat{j}}}, LCS}, & 0 < \psi \leq 1 \end{cases},$$

$$\bar{\varepsilon}^{x_{L_{\hat{j}}},LCS}_{L_{\hat{j}},1} \approx 1 + \chi_{L_{\hat{j}}} \sqrt{N_{L_{\hat{j}}}} \sum_{p=1}^{a_{L,II}} \sum_{\Delta_L=p} \frac{\Phi_L c_{L,II} \hat{\omega}_L^{-\varphi_L-1}}{\left(\alpha_{L_{\hat{j}}}\gamma_0\right)^{\varphi_L}} \Xi_{L,4},$$

and

$$\bar{\varepsilon}^{x_{L_{\hat{j}}},LCS}_{L_{\hat{j}},2} \approx 1 + \frac{\chi_{L_{\hat{j}}} \alpha_{L_{\hat{j}}} \sqrt{N_{L_{\hat{j}}}}}{\gamma_0 \psi^2 \alpha^2_{H_{\hat{i}}}} \sum_{p=1}^{a_{L,II}} \sum_{\Delta_L=p} \sum_{q=0}^{\varphi_L} \begin{pmatrix} \varphi_L \\ q \end{pmatrix}$$
$$\times \left(-\frac{1}{\gamma_0 \psi \alpha_{H_{\hat{i}}}}\right)^q \Phi_L c_{L,II} e^{\frac{\omega_L}{\gamma_0 \eta \alpha_{H_{\hat{i}}}}} \mathcal{A}_L\left(\phi_{L_{\hat{j}}}, \kappa_{L_{\hat{j}}}\right).$$

Proof: The proof of this theorem can be carried out in the same way as the proof of Theorem 2, where (29) is used instead of (26). ∎

IV. PROPOSED ANALYTICAL FRAMEWORK FOR OPTIMAL POWER ALLOCATION AND MINIMUM BLOCKLENGTH

By following the average BLER analysis presented in Section III, this section provides the derivation of the optimal power allocation coefficients for a minimum blocklength[5] based on asymptotic average BLER in high SNR regime, and it also presents the analytical comparison of the minimum blocklength of NOMA with the OMA case.

A. Asymptotic Average BLER Analysis

As discussed in [29, 30], the average BLER, $\bar{\varepsilon}^{x_V}_U$, in (22) can be simplified by utilizing the first-order Riemann integral approximation, i.e., $\int_a^b f(x)dx = (b-a)f\left(\frac{a+b}{2}\right)$, as follows:

$$\bar{\varepsilon}^{x_V}_U \approx \chi_V \sqrt{N_V} (\mu_V - v_V) F_{\gamma^{x_V}_U}\left(\frac{v_V + \mu_V}{2}\right). \quad (34)$$

By substituting v_V and μ_V defined in (20) into (34), $\bar{\varepsilon}^{x_V}_U$ is rewritten as

$$\bar{\epsilon}^{x_V}_U \approx F_{\gamma^{x_V}_U}(\beta_V), \quad (35)$$

where β_V is defined in (21).

By using the series representation of e^x in [53, Eq. 1.211], i.e., $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$, the asymptotic CDF of $\gamma^{x_{H_{\hat{i}}}}_{H_{\hat{i}}}$, $\gamma^{x_{H_{\hat{i}}}}_{L_{\hat{j}}}$, and $\gamma^{x_{L_{\hat{j}}}}_{L_{\hat{j}}}$ are respectively given by

$$F^{s,\infty}_{\gamma^{x_{H_{\hat{i}}}}_{H_{\hat{i}}}}(x) = F^{s,\infty}_{g_{SH}}(B_x) \stackrel{\gamma_0\to\infty}{\approx} \frac{(m_H B_x)^{b_H a_{H,r}}}{(b_H!)^{a_{H,r}} \lambda^{b_H a_{H,r}}_{SH}}, \quad (36)$$

$$F^{s,\infty}_{\gamma^{x_{H_{\hat{i}}}}_{L_{\hat{j}}}}(x) \stackrel{\gamma_0\to\infty}{\approx} \frac{(m_L B_x)^{b_L a_{L,r}}}{(b_L!)^{a_{L,r}} \lambda^{b_L a_{L,r}}_{SL}}, \quad (37)$$

and

$$F^{s,\infty}_{\gamma^{x_{L_{\hat{j}}}}_{L_{\hat{j}}}}(x) \stackrel{\gamma_0\to\infty}{\approx} \frac{(m_L \hat{B}_x)^{b_L a_{L,r}}}{(b_L!)^{a_{L,r}} \lambda^{b_L a_{L,r}}_{SL}}, \quad (38)$$

where $s \in \{HCS, LCS\}$, $r = \begin{cases} I, & \text{if } s = HCS \\ II, & \text{if } s = LCS \end{cases}$, $B_x = \frac{x}{\gamma_0\left(\alpha_{H_{\hat{i}}} - \alpha_{L_{\hat{j}}}x\right)}$, and $\hat{B}_x = \frac{x}{\gamma_0\left(\alpha_{L_{\hat{j}}} - \psi\alpha_{H_{\hat{i}}}x\right)}$. From (34) - (38), the asymptotic average BLER at users $H_{\hat{i}}$ and $L_{\hat{j}}$ are respectively expressed as

$$\bar{\varepsilon}^{s,\infty}_{H_{\hat{i}}} \approx \frac{\left(m_H B_{\beta_{H_{\hat{i}}}}\right)^{b_H a_{H,r}}}{(b_H!)^{a_{H,r}} \lambda^{b_H a_{H,r}}_{SH}}, \quad (39)$$

and

$$\bar{\varepsilon}^{s,\infty}_{L_{\hat{j}}} = \bar{\varepsilon}^{x_{H_{\hat{i}}},s}_{L_{\hat{j}},\infty} + \left(1 - \bar{\varepsilon}^{x_{H_{\hat{i}}},s}_{L_{\hat{j}},\infty}\right) \bar{\varepsilon}^{x_{L_{\hat{j}}},s}_{L_{\hat{j}},\infty}$$
$$\approx \bar{\varepsilon}^{x_{H_{\hat{i}}},s}_{L_{\hat{j}},\infty} + \bar{\varepsilon}^{x_{L_{\hat{j}}},s}_{L_{\hat{j}},\infty}$$
$$\approx \frac{\left(m_L B_{\beta_{H_{\hat{i}}}}\right)^{b_L a_{L,r}}}{(b_L!)^{a_{L,r}} \lambda^{b_L a_{L,r}}_{SL}} + \frac{\left(m_L \hat{B}_{\beta_{L_{\hat{j}}}}\right)^{b_L a_{L,r}}}{(b_L!)^{a_{L,r}} \lambda^{b_L a_{L,r}}_{SL}}. \quad (40)$$

From (39) and (40), the diversity order at users $H_{\hat{i}}$ and $L_{\hat{j}}$ are respectively given by [45]:

$$D_{H_{\hat{i}}} = -\lim_{\gamma_0\to\infty} \frac{\log\left(\bar{\varepsilon}^{s,\infty}_{H_{\hat{i}}}\right)}{\log(\gamma_0)}$$
$$= \begin{cases} m_H K_S K_H I, & \text{for HCS method} \\ m_H K_H I, & \text{for LCS method} \end{cases}, \quad (41)$$

and

$$D_{L_{\hat{j}}} = -\lim_{\gamma_0\to\infty} \frac{\log\left(\bar{\varepsilon}^{s,\infty}_{L_{\hat{j}}}\right)}{\log(\gamma_0)}$$
$$= \begin{cases} m_L K_L J, & \text{for HCS method} \\ m_L K_S K_L J, & \text{for LCS method} \end{cases}. \quad (42)$$

Remark 1. *For both TAS/SC and TAS/MRC schemes, the diversity orders at users $H_{\hat{i}}$ and $L_{\hat{j}}$, denoted by $\left(D_{H_{\hat{i}}}, D_{L_{\hat{j}}}\right)$, are $(m_H K_S K_H I, m_L K_L J)$ for HCS method, and $(m_H K_H I, m_L K_S K_L J)$ for LCS method. This reveals that the users $H_{\hat{i}}$ and $L_{\hat{j}}$ have achieved full diversity order with HCS and LCS methods, respectively. Furthermore, the system performance of user $H_{\hat{i}}$ can be improved by increasing m_H, K_S, K_H, and I with HCS method, and by increasing m_H, K_H, and I with LCS method. Meanwhile, the growth of m_L, K_L, and J with HCS method, and m_L, K_S, K_L, and J with LCS method can help enhancing the system performance of user $L_{\hat{j}}$.*

B. Power and Blocklength Optimization at High SNR

In this subsection, we focus on the problem of blocklength minimization[6] subject to BLER targets and power allocation coefficients to guarantee the reliability requirement and reduce the transmission latency for SPC-based MIMO NOMA systems [7, 29, 36]. To determine the values of power allocation coefficients (i.e., $\alpha_{H_{\hat{i}}}$ and $\alpha_{L_{\hat{j}}}$) at which a minimum

---

[5]In this paper, we focus on minimizing the blocklength in NOMA-based SPC systems to reduce the latency for two users having the best channel conditions, which are selected from two predefined clusters H and L. However, investigating suitable user pairing methods to guarantee latency requirements for the users having weak channel gains is an important research issue to be addressed in future works.

[6]It is noted that addressing the optimization problems subject to latency requirement (e.g., decoding error probability minimization under the latency constraint) is also an important issue to be tackled in the SPC-based systems to ensure the expected latency [54, 55]. This would be a noteworthy problem to investigate in future work.

blocklength N_U $\left(U \in \left\{ H_{\hat{i}}, L_{\hat{j}} \right\} \right)$ is achieved to guarantee the reliability target $\bar{\varepsilon}_U^{th}$, the following problem needs to be addressed

$$\min_{\alpha_{H_{\hat{i}}}, \alpha_{L_{\hat{j}}}} \quad N_U \tag{43a}$$

$$\text{s.t.} \quad \bar{\varepsilon}_U \leq \bar{\varepsilon}_U^{th}, \tag{43b}$$

$$\alpha_{H_{\hat{i}}} + \alpha_{L_{\hat{j}}} = 1, \ 0 < \alpha_{L_{\hat{j}}} < 0.5, \tag{43c}$$

where the constraint (43c) is obtained based on the NOMA principle presented in Section II, in which $0 < \alpha_{L_{\hat{j}}} < \alpha_{H_{\hat{i}}}$ and $\alpha_{L_{\hat{j}}} + \alpha_{H_{\hat{i}}} = 1$, leading to $\alpha_{L_{\hat{j}}} < 0.5$. It is noted that $\alpha_{H_{\hat{i}}} = 1 - \alpha_{L_{\hat{j}}}$ and $\bar{\varepsilon}_U$ is a decreasing function of N_U. The problem in (43) can be simplified as

$$\min_{\alpha_{L_{\hat{j}}}} \quad N_U \tag{44a}$$

$$\text{s.t.} \quad \bar{\varepsilon}_U = \bar{\varepsilon}_U^{th}, \tag{44b}$$

$$0 < \alpha_{L_{\hat{j}}} < 0.5. \tag{44c}$$

By substituting (39) into (44b) for user $H_{\hat{i}}$ and (40) into (44b) for user $L_{\hat{j}}$, the blocklengths of users $H_{\hat{i}}$ and $L_{\hat{j}}$ with s ($s \in \{HCS, LCS\}$) method are respectively calculated as

$$N_{H_{\hat{i}},s} = \frac{n_{H_{\hat{i}}}}{\log_2 \left(\frac{1+\tau_{H,r}}{1+\alpha_{L_{\hat{j}}}\tau_{H,r}} \right)}, \tag{45}$$

and

$$N_{L_{\hat{j}},s} = \frac{n_{L_{\hat{j}}}}{\log_2 \left\{ 1 + \frac{\alpha_{L_{\hat{j}}}\tau_{L,r}}{1+\psi\left(1-\alpha_{L_{\hat{j}}}\right)\tau_{L,r}} \right\}}, \tag{46}$$

where $\eta_{H,r} = \frac{m_H^{b_H a_{H,r}}}{(b_H!)^{a_{H,r}} \lambda_{SH}^{b_H a_{H,r}} \gamma_0^{b_H a_{H,r}}}$, $\eta_{L,r} = \frac{(b_H!)^{\frac{b_L a_{L,r}}{b_H}}}{(b_L!)^{a_{L,r}}} \left(\frac{m_L \lambda_{SH}}{m_H \lambda_{SL}} \right)^{b_L a_{L,r}}$, $\hat{\eta}_{L,r} = \frac{m_L^{b_L a_{L,r}}}{(b_L!)^{a_{L,r}} \lambda_{SL}^{b_L a_{L,r}}}$, $\tau_{H,r} = \left(\bar{\varepsilon}_{H_{\hat{i}},r}^{th}/\eta_{H,r} \right)^{1/b_H a_{H,r}}$, and $\tau_{L,r} = \gamma_0 \left[\frac{\bar{\varepsilon}_{L_{\hat{j}},r}^{th} - \eta_{L,r}\left(\varepsilon_{H_{\hat{i}},r}^{th}\right)^{\frac{b_L a_{L,r}}{b_H a_{H,r}}}}{\hat{\eta}_{L,r}} \right]^{1/b_L a_{L,r}}$.

From (45) and (46), the derivative of $N_{H_{\hat{i}},s}$ and $N_{L_{\hat{j}},s}$ with respect to $\alpha_{L_{\hat{j}}}$ are derived as

$$\frac{\partial N_{H_{\hat{i}},s}}{\partial \alpha_{L_{\hat{j}}}} = \frac{n_{H_{\hat{i}}}\tau_{H,r}}{\left(1+\alpha_{L_{\hat{j}}}\tau_{H,r}\right)\left[\log_2\left(\frac{1+\tau_{H,r}}{1+\alpha_{L_{\hat{j}}}\tau_{H,r}}\right)\right]^2 \ln 2} > 0, \tag{47}$$

and

$$\frac{\partial N_{L_{\hat{j}},s}}{\partial \alpha_{L_{\hat{j}}}} = -\frac{n_{L_{\hat{j}}}\left(1+\psi\tau_{L,r}\right)\hat{\tau}_{L,r}}{\left(1+\alpha_{L_{\hat{j}}}\hat{\tau}_{L,r}\right)\left[\log_2\left(1+\alpha_{L_{\hat{j}}}\hat{\tau}_{L,r}\right)\right]^2 \ln 2} < 0, \tag{48}$$

where $\hat{\tau}_{L,r} = \frac{\tau_{L,r}}{1+\psi\left(1-\alpha_{L_{\hat{j}}}\right)\tau_{L,r}}$. Thus, $N_{H_{\hat{i}},s}$ is an increasing function of $\alpha_{L_{\hat{j}}}$, whereas $N_{L_{\hat{j}},s}$ is a decreasing function of $\alpha_{L_{\hat{j}}}$. Therefore, to guarantee both reliability targets $\bar{\varepsilon}_{H_{\hat{i}},r}^{th}$ and $\bar{\varepsilon}_{L_{\hat{j}},r}^{th}$, the minimum blocklength is obtained by addressing

---

Algorithm 1: Proposed Power Allocation Algorithm for SPC-Based MIMO NOMA System

Data : $n_{H_{\hat{i}}}$, $n_{L_{\hat{j}}}$, γ_0, $\bar{\varepsilon}_{H_{\hat{i}},r}^{th}$, $\bar{\varepsilon}_{L_{\hat{j}},r}^{th}$, K_S, K_H, K_L, I, J, λ_{SH}, λ_{SL}, and tolerance μ.

Result: Determine optimal power allocation coefficient $\alpha_{L_{\hat{j}},opt}$.

1 Initialize: $\alpha_{L_{\hat{j}}}^- \leftarrow 0$, $\alpha_{L_{\hat{j}}}^+ \leftarrow 0.5$, and $\hat{\alpha}_{L_{\hat{j}}} \leftarrow \frac{\alpha_{L_{\hat{j}}}^- + \alpha_{L_{\hat{j}}}^+}{2}$;

2 **while** $\left| f\left(\hat{\alpha}_{L_{\hat{j}}}\right) \right| > \mu$ **do**

3   **if** $f\left(\hat{\alpha}_{L_{\hat{j}}}\right) f\left(\alpha_{L_{\hat{j}}}^-\right) > 0$ **then**

4   Set $\alpha_{L_{\hat{j}}}^- \leftarrow \hat{\alpha}_{L_{\hat{j}}}$;

5   **else**

6   Set $\alpha_{L_{\hat{j}}}^+ \leftarrow \hat{\alpha}_{L_{\hat{j}}}$;

7   **end**

8   Set $\hat{\alpha}_{L_{\hat{j}}} \leftarrow \frac{\alpha_{L_{\hat{j}}}^- + \alpha_{L_{\hat{j}}}^+}{2}$ and compute $f\left(\hat{\alpha}_{L_{\hat{j}}}\right)$ based on (45) and (46);

9 **end**

10 Set $\alpha_{L_{\hat{j}},opt} \leftarrow \hat{\alpha}_{L_{\hat{j}}}$;

11 **Return** $\alpha_{L_{\hat{j}},opt}$;

$N_{H_{\hat{i}},s} = N_{L_{\hat{j}},s} = N_{opt,s}$ and the problem of minimizing blocklength in (44) is rewritten as

$$\min_{\alpha_{L_{\hat{j}}}} \quad N_{opt,s} \tag{49a}$$

$$\text{s.t.} \quad \bar{\varepsilon}_{H_{\hat{i}}}^s = \bar{\varepsilon}_{H_{\hat{i}},r}^{th}, \tag{49b}$$

$$\bar{\varepsilon}_{L_{\hat{j}},s} = \bar{\varepsilon}_{L_{\hat{j}},r}^{th}, \tag{49c}$$

$$0 < \alpha_{L_{\hat{j}}} < 0.5. \tag{49d}$$

Given this context, the optimal power allocation coefficient $\alpha_{L_{\hat{j}},opt}$ to minimize $N_{opt,s}$ can be achieved by solving the equation $f\left(\alpha_{L_{\hat{j}}}\right) = N_{L_{\hat{j}},s} - N_{H_{\hat{i}},s} = 0$, which is addressed in Algorithm 1. The minimum blocklength $N_{opt,r}$ is attained by substituting $\alpha_{L_{\hat{j}},opt}$ into (45) as follows:

$$N_{opt,s} = \frac{n_{H_{\hat{i}}}}{\log_2 \left(\frac{1+\left(\bar{\varepsilon}_{H_{\hat{i}},r}^{th}/\eta_{H,r}\right)^{1/b_H a_{H,r}}}{1+\alpha_{L_{\hat{j}},opt}\left(\bar{\varepsilon}_{H_{\hat{i}},r}^{th}/\eta_{H,r}\right)^{1/b_H a_{H,r}}} \right)}. \tag{50}$$

C. Comparison of MIMO NOMA and MIMO OMA Schemes

To perform the comparison between MIMO NOMA and MIMO OMA schemes, we consider a scenario where users $H_{\hat{i}}$ and $L_{\hat{j}}$ are served simultaneously by using NOMA or over different time-slots by utilizing OMA (i.e., time division multiple access). Herein, a MIMO scheme, i.e., TAS, is utilized for both NOMA and OMA scenarios to reduce the complexity of the signal processing and feedback overhead [26, 27]. With OMA transmission, the minimum blocklength, $N_{OMA,s}$ ($r \in \{I, II\}$), is the summation of the minimum blocklengths for users $H_{\hat{i}}$ and $L_{\hat{j}}$, $\hat{N}_{H_{\hat{i}}}$ and $\hat{N}_{L_{\hat{j}}}$. Similar to

the derivation of blocklengths for users $H_{\hat{i}}$ and $L_{\hat{j}}$ in Section IV-B, $N_{OMA,s}$ in the high SNR regime is calculated as

$$N_{OMA,s} = \hat{N}_{H_{\hat{i}}} + \hat{N}_{L_{\hat{j}}}$$
$$= \frac{n_{H_{\hat{i}}}}{\log_2\left[1 + \left(\bar{\varepsilon}^{th}_{H_{\hat{i}},r}/\eta_{H,r}\right)^{1/b_H a_{H,r}}\right]} \tag{51}$$
$$+ \frac{n_{L_{\hat{j}}}}{\log_2\left[1 + \gamma_0\left(\bar{\varepsilon}^{th}_{L_{\hat{j}},r}/\hat{\eta}_{L,r}\right)^{1/b_L a_{L,r}}\right]}.$$

From (50) and (51), the blocklength gap between NOMA and OMA, ΔN_s, is given by

$$\Delta N_r = N_{OMA,r} - N_{opt,r} \approx \hat{N}_{H_{\hat{i}},r} > 0. \tag{52}$$

Thus, OMA transmission needs a longer blocklength than NOMA transmission to serve the users $H_{\hat{i}}$ and $L_{\hat{j}}$.

V. NUMERICAL RESULTS

In this section, we provide numerical results in terms of average BLER and minimum blocklength to characterize the effects of the proposed protocols, i.e., HCS and LCS methods with TAS/SC and TAS/MRC schemes discussed in Section II-A, on the system performance in designing an SPC-based MIMO NOMA network. It is noted that the analysis of these performance metrics have practical significance for the reliability and latency performance evaluation of wireless systems [29–31, 34–36]. The predetermined simulation parameters are set as follows [29–31]: the number of information bits $n_{H_{\hat{i}}} = n_{L_{\hat{j}}} = 80$ bits; the blocklength $N_{H_{\hat{i}}} = N_{L_{\hat{j}}} = 100$; the path loss exponent $\theta = 2.5$; the distances $d_{SH} = d_{SL} = 5$ (m); the power allocation coefficients $\alpha_{H_{\hat{i}}} = 0.7$, and $\alpha_{H_{\hat{i}}} = 0.3$; the reliability targets $\bar{\varepsilon}^{th}_{H_{\hat{i}}} = 10^{-7}$ and $\bar{\varepsilon}^{th}_{L_{\hat{j}}} = 10^{-6}$.

To evaluate our BLER performance analysis carried out in Section III, we provide the numerical outcomes through the following three types of result: i) Analytical result (Ana.), ii) Asymptotic result (Asymp.), and iii) Simulation result (Sim.). For the simulation results, similar to the method used in [29–31, 34, 36], we create 10^6 Nakagami-m channel realizations generated randomly through the Nakagami-m distribution given in (1) and (2) for all the considered schemes. The respective average BLERs are then computed by averaging the instantaneous BLERs according to these generated channel realizations while considering the Gaussian-coded symbols instead of the real symbol constellations. The definition of the instantaneous BLER is provided in (19), which has also been used in [29–31, 34, 36]. Furthermore, the different values of Nakagami-m fading parameters, i.e., m_H and m_L, are considered in the presented numerical results. For the analytical and asymptotic results, they are obtained by adopting the expressions derived in (30), (31), (32), (33) for the analytical results, and (39), and (40) for the asymptotic results, respectively.

In Fig. 2, we perform the rate comparison of SPC and long-packet communications (LPC) for user $H_{\hat{i}}$ to gain more insights on SPC. Note that based on (17) and the normal approximation method in [7], the achievable rate of user $H_{\hat{i}}$ with SPC can be approximated as: $R^{SPC}_{H_{\hat{i}}} \approx \log_2\left(1 + \gamma^{x_{H_{\hat{i}}}}_{H_{\hat{i}}}\right) -$
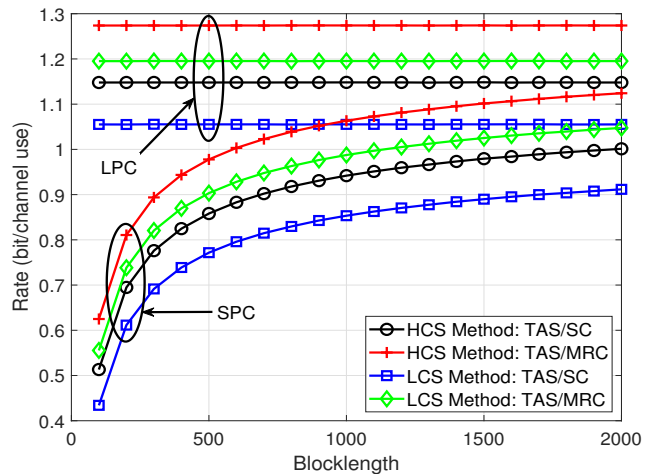

Fig. 2. Rate comparison between SPC and LPC, where $\gamma_0 = 20$ (dB), $m_H = m_L = 2$, and $K_S = K_H = K_L = I = J = 2$.

$\sqrt{v/N}Q^{-1}(\varepsilon) + \log_2 N/2N$. We can observe from this figure that when the blocklength of user $H_{\hat{i}}$ ($N_{H_{\hat{i}}}$) increases, the achievable rate of this user with LPC ($R^{LPC}_{H_{\hat{i}}}$) is unchanged, whereas the rate with the SPC, $R^{SPC}_{H_{\hat{i}}}$ grows up. This can be explained by the fact that the LPC is implemented under the assumption of infinite blocklength to obtain the Shannon's channel rate, i.e., $\log_2\left(1 + \gamma^{x_{H_{\hat{i}}}}_{H_{\hat{i}}}\right)$, which is not influenced by $N_{H_{\hat{i}}}$. In contrast, finite blocklength is utilized in the SPC scenario. Given the above approximation of $R^{SPC}_{H_{\hat{i}}}$, the increase in $N_{H_{\hat{i}}}$ leads to the higher value of $R^{SPC}_{H_{\hat{i}}}$ in this case, and $R^{SPC}_{H_{\hat{i}}} \to R^{LPC}_{H_{\hat{i}}}$ when $N_{H_{\hat{i}}} \to \infty$. Since (17) is applied for both users $H_{\hat{i}}$ and $L_{\hat{j}}$, the same conclusions can also be achieved for user $L_{\hat{j}}$. Furthermore, this figure indicates that HCS method with TAS/MRC provides the best performance for user $H_{\hat{i}}$. The effects of the proposed protocols on the performance of users $H_{\hat{i}}$ and $L_{\hat{j}}$ will be clearly analyzed in the following.

Similar to [29–31], we utilize the simulation results to evaluate the correctness of our analysis. Specifically, in Figs. 3 and 4, we plot the average BLERs at users $H_{\hat{i}}$ and $L_{\hat{j}}$ as a function of γ_0 with different methods (i.e., HCS method with TAS/SC or TAS/MRC and LCS method with TAS/SC or TAS/MRC). As can be observed from these figures, the BLER gap between the approximated analytical results and the simulation results are very small. Furthermore, the asymptotic curves accurately predict the system performance trend in the higher γ_0 regime. These observations confirm the correctness of our analysis in Section III. In addition, Figs. 3 and 4 show that HCS method achieves better performance (i.e., lower value of average BLER is observed) for user $H_{\hat{i}}$ over LCS method, whereas LCS method outperforms HCS method in terms of the system performance for user $L_{\hat{j}}$. This result is achieved based on the fact that HCS and LCS methods are proposed to improve the received signal quality at users $H_{\hat{i}}$ and $L_{\hat{j}}$, respectively, as discussed in Section II-A. Furthermore, these figures indicate that TAS/MRC scheme is better than TAS/SC in improving the system performance.

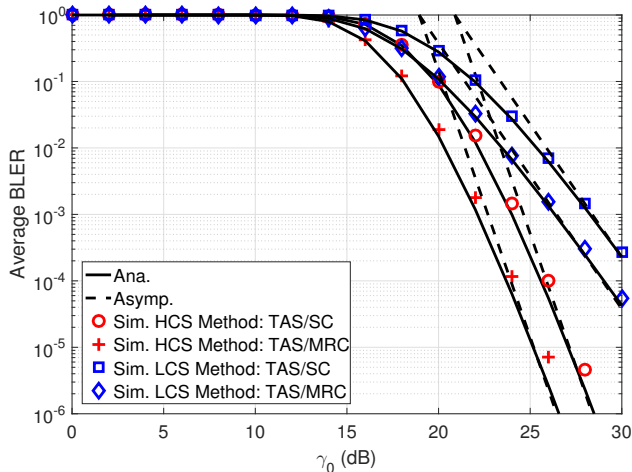In Figs. 5 and 6, we investigate the effects of the number

Fig. 3. Average BLER at user $H_{\hat{i}}$ vs. γ_0 with different methods, where $m_H = m_L = 2$ and $(K_S, K_H, I) = (2, 2, 1)$.
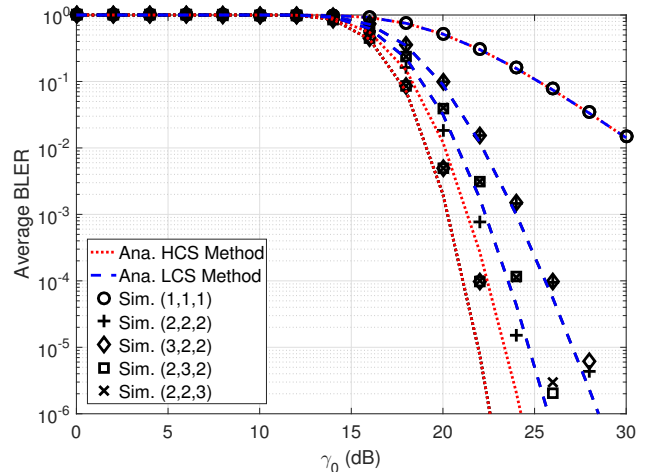

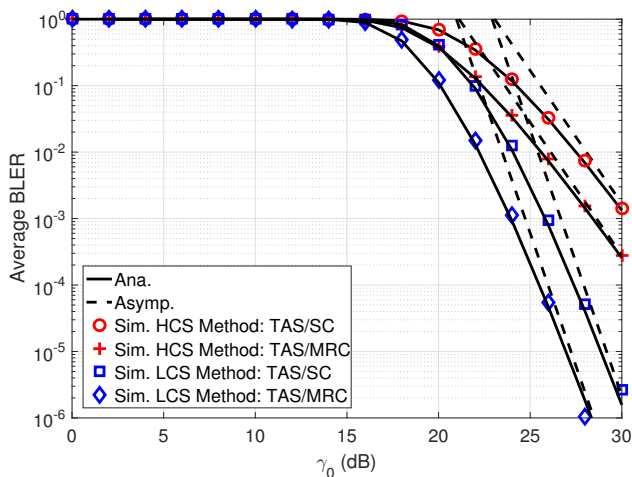Fig. 5. Average BLER at user $H_{\hat{i}}$ vs. γ_0 with different values of (K_S, K_H, I), where $m_H = m_L = 2$.

Fig. 4. Average BLER at user $L_{\hat{j}}$ vs. γ_0 with different methods, where $\psi = 0$, $m_H = m_L = 2$ and $(K_S, K_L, J) = (2, 2, 1)$.
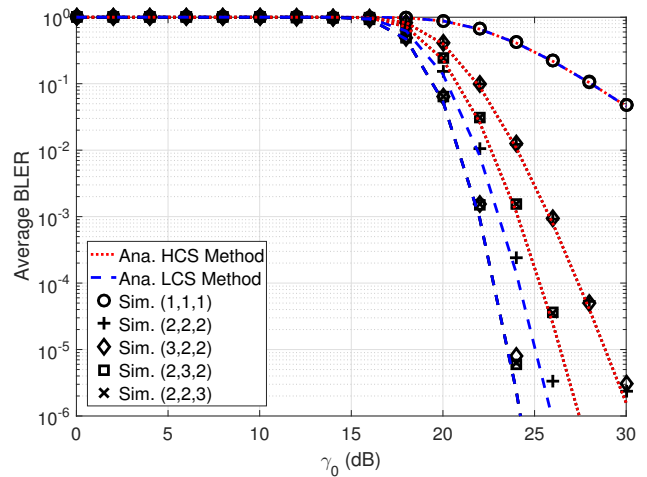

Fig. 6. Average BLER at user $L_{\hat{j}}$ vs. γ_0 with different values of (K_S, K_L, J), where $\psi = 0$ and $m_H = m_L = 2$.

of users at clusters H (I) and L (J), and the number of antennas at BS S (K_S), users $H_{\hat{i}}$ (K_H), and $L_{\hat{j}}$ (K_L), on the system performance. Specifically, Fig. 5 shows the variation of average BLER at user $H_{\hat{i}}$ with respect to γ_0 with different values of K_S, K_H, and I, denoted by (K_S, K_H, I), in case of utilizing HCS and LCS methods with the TAS/SC scheme. Meanwhile, Fig. 6 plots the average BLER at user $L_{\hat{j}}$ versus γ_0 with different values of (K_S, K_L, J) when using HCS and LCS methods with the TAS/SC scheme. These two figures indicate that as K_S, K_H, K_L, I, and J are all equal to one, HCS and LCS methods result in the same curves. Furthermore, the system performance can be significantly improved by increasing (K_S, K_H, I) for user $H_{\hat{i}}$ and (K_S, K_L, J) for user $L_{\hat{j}}$. It is noted that the variation of K_S in LCS method does not impact the system performance at user $H_{\hat{i}}$ (see Fig. 5). The same conclusion can be derived for user $L_{\hat{j}}$ when observing the change of K_S in HCS method (see Fig. 6). The reason for this is based on the nature of HCS and LCS methods as

mentioned in Section II-A and the discussion part of Figs. 3 and 4. This phenomenon also confirms our analysis of diversity order for users $H_{\hat{i}}$ and $L_{\hat{j}}$, as shown in Section IV-A.

Fig. 7 depicts the average BLER at user $L_{\hat{j}}$ as a function of γ_0 with different values of the residual interference level caused by the ISIC, i.e., ψ, in case of using HCS and LCS methods with TAS/SC scheme. In other results presented in this section, we investigate a scenario where the value of ψ is fixed to evaluate the effects of other parameters such as antenna-user selection methods, the number of users, the number of antennas, fading parameters, and power allocation coefficients, on the system performance. In contrast, Fig. 7 shows how the variation of ψ affects the BLER performance of SPC in the considered MIMO NOMA system. We can observe from this figure that the increase in ψ leads to the higher interference as in (15), making the system performance of user $L_{\hat{j}}$ lower. Thus, user $L_{\hat{j}}$ can achieve the best performance when $\psi = 0$, where the perfect SIC is observed, which may
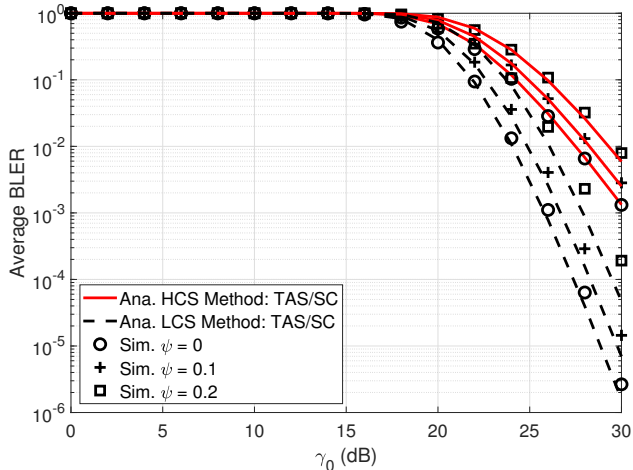
Fig. 7. Average BLER at user $L_{\hat{j}}$ vs. γ_0 with different values of ψ (residual interference caused due to the ISIC process), where $m_H = m_L = 2$ and $(K_S, K_L, J) = (2, 2, 1)$.

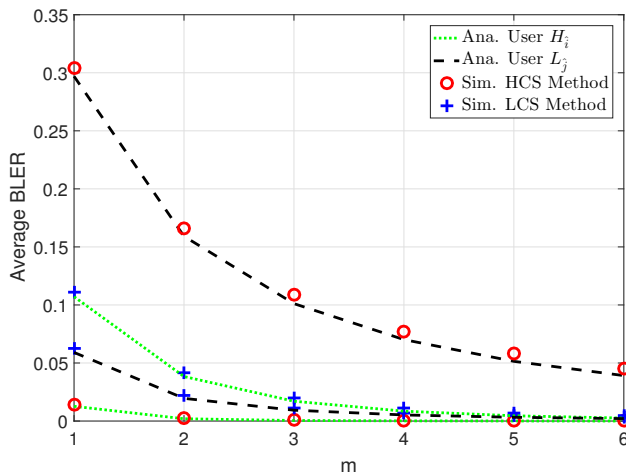Fig. 9. Minimum blocklength for users $H_{\hat{i}}$ and $L_{\hat{j}}$ vs. $\alpha_{L_{\hat{j}}}$ with different methods, where $\psi = 0$, $m_H = m_L = 2$, $K_S = K_H = K_L = I = J = 2$, and $\gamma_0 = 20$ (dB).

Fig. 8. Average BLER at users $H_{\hat{i}}$ and $L_{\hat{j}}$ vs. m with different methods, where $\gamma_0 = 20$ (dB) and $(K_S, K_H, K_L, I, J) = (2, 2, 2, 1, 1)$.

Fig. 10. Blocklength comparison between NOMA and OMA.

be difficult to obtain in practical scenarios.

In Fig. 8, we consider the change of average BLER at users $H_{\hat{i}}$ and $L_{\hat{j}}$ with respect to the fading parameters, i.e., m_H and m_L, in case of using HCS and LCS methods with the TAS/SC scheme. Herein, we set $m_H = m_L = m$. Given the considered systems, it is noteworthy that m_H and m_L are independent and only affect users $H_{\hat{i}}$ and $L_{\hat{j}}$, respectively. We can see from this figure that the performance of users $H_{\hat{i}}$ and $L_{\hat{j}}$ can be improved with the increase in m_H and m_L, respectively, due to the better channel quality. More precisely, when $m = 1$, Nakagami-m fading corresponds to Rayleigh fading and the worst performance can be observed. In case of $m = (K+1)^2/(2K+1)$, it approximates the Rician fading with parameter K [41]. This result also verifies the diversity order outcomes obtained in (41) and (42). Furthermore, similar to Figs. 3, 4, 5, and 6, Fig. 8 indicates that HCS and LCS methods provide the best performance for users $H_{\hat{i}}$ and $L_{\hat{j}}$, respectively.
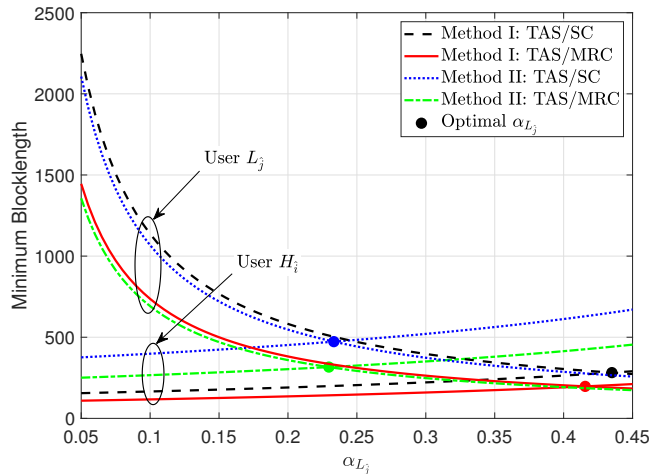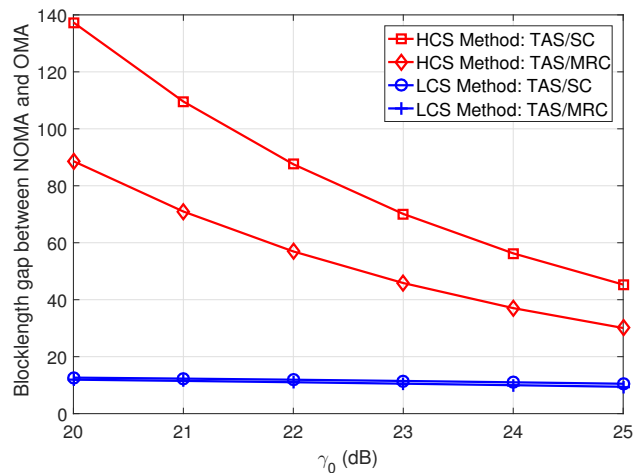
Fig. 9 depicts the effect of power allocation coefficient $\alpha_{L_{\hat{j}}}$ on the blocklength of users $H_{\hat{i}}$ $\left(N_{H_{\hat{i}}}\right)$ and $L_{\hat{j}}$ $\left(N_{L_{\hat{j}}}\right)$. One can see from this figure that $N_{H_{\hat{i}}}$ and $N_{L_{\hat{j}}}$ are increasing and decreasing functions of $\alpha_{L_{\hat{j}}}$, respectively. Thus, there exists an optimal value of $\alpha_{L_{\hat{j}}}$, at which the minimum blocklength for both users $H_{\hat{i}}$ and $L_{\hat{j}}$ is achieved. The value of optimal $\alpha_{L_{\hat{j}}}$ for different cases (i.e., HCS method with TAS/SC or TAS/MRC; LCS method with TAS/SC or TAS/MRC) can be found out by using Algorithm 1 and then the minimum blocklength is calculated by using (50).

In Fig. 10, we perform the minimum blocklength comparison between NOMA and OMA transmissions (N_{opt} and N_{OMA}) to clarify the benefits of NOMA over OMA in short-packet transmissions. As can be seen from this figure, the higher blocklength gap between NOMA and OMA, i.e., Δ_N (calculated from (52)), is achieved in case of using HCS method and TAS/SC scheme. This implies that the benefits of MIMO NOMA versus MIMO OMA in terms of minimum

blocklength are more pronounced when utilizing HCS method as compared to LCS method. Furthermore, Δ_N is positive, hence, N_{opt} is always smaller than N_{OMA}. In other words, MIMO NOMA can lower the transmission latency of SPC systems as compared to the MIMO OMA case.

From the above achieved results, we provide some useful insights when considering SPC in the considered MIMO NOMA system as follows: i) Compared to LPC, SPC can fulfill more stringent requirements of reliability and latency for MIMO NOMA but achieves lower rate performance; ii) Transmission latency of MIMO NOMA is smaller than that of MIMO OMA in SPC scenario; and iii) Minimum blocklength for MIMO NOMA is achieved at a certain value of power allocation coefficients such that blocklengths of NOMA users are the same.

VI. CONCLUSIONS

In this paper, we analyzed the performance of SPC in a QoS-based multiuser downlink MIMO NOMA system over a Nakagami-*m* fading channel under the ISIC scenario in terms of the average BLER and minimum blocklength. Specifically, we considered the user pairing to perform NOMA, where users are selected from two user clusters having different priority levels. Furthermore, we investigated different MIMO schemes including TAS for BS, SC and MRC for users, and proposed two antenna-user selection methods, i.e., HCS and LCS to design effective communication protocols for the SPC-based MIMO NOMA systems. We characterized the system performance by deriving the approximate and asymptotic (in the high SNR regime) closed-form expressions of the average BLER at the users. From the asymptotic average BLER, we carried out an analysis of diversity order, minimum block-length, and optimal power allocation. The analytical results verified by simulation results indicated that the system performance decreases with the increase in the value of the residual interference caused by the ISIC process. In addition, among the proposed schemes, the HCS method with TAS/MRC and the LCS method with TAS/MRC provide the best performance with full diversity gains for the users selected from the high-priority and low-priority user clusters, respectively. Moreover, it has been demonstrated that MIMO can significantly improve the performance of NOMA systems with short-packets, and MIMO NOMA outperforms MIMO OMA in ensuring low-latency transmissions.

APPENDIX A
PROOF OF PROPOSITION 1

Using (3) and (5), the CDF of g_{SH} in this case is given by [37]

$$F_{g_{SH}}^{HCS}(x) = \left(1 - \sum_{p=0}^{b_H-1} \frac{m_H^p}{p!\lambda_{SH}^p} x^p e^{-\frac{m_H x}{\lambda_{SH}}}\right)^{a_{H,I}}. \tag{53}$$

Applying binomial expansion in [53, Eq. (1.111)], (53) can be rewritten as

$$F_{g_{SH}}^{HCS}(x) = 1 + \sum_{p=1}^{a_{H,I}} \phi \underbrace{\left(\sum_{q=0}^{b_H-1} \frac{m_H^q x^q}{q!\lambda_{SH}^q}\right)^p}_{\Psi}, \tag{54}$$

where $\phi = \binom{a_{H,I}}{p}(-1)^p e^{-\frac{pm_H x}{\lambda_{SH}}}$.

To derive (54), we first resolve Ψ in (54) by utilizing the multinomial theorem as follows:

$$\Psi = \sum_{\Delta_H=p} \psi \left[\prod_{q=0}^{b_H-1}\left(\frac{m_H^q}{q!\lambda_{SH}^q}\right)^{\delta_{H,q}}\right] x^{\varphi_H}, \tag{55}$$

where $\psi = \binom{p}{\delta_{H,0},\ldots,\delta_{H,b_H-1}}$.

The final expression of $F_{g_{SH}}^{HCS}(x)$ is achieved as in (25) by substituting (55) into (54).

APPENDIX B
PROOF OF THEOREM 1

From (12), the CDF of $\gamma_{H_{\hat{i}}}^{x_{H_{\hat{i}}}}$ is given by

$$F_{\gamma_{H_{\hat{i}}}^{x_{H_{\hat{i}}}}}(x) = \Pr\left\{\frac{\alpha_{H_{\hat{i}}}\gamma_0 g_{SH}}{\alpha_{L_{\hat{j}}}\gamma_0 g_{SH}+1} < x\right\} \\ = F_{g_{SH}}(B_x), \tag{56}$$

where (56) is obtained under the condition $\alpha_{H_{\hat{i}}} - \alpha_{L_{\hat{j}}} x > 0$ and $B_x = \frac{x}{\gamma_0\left(\alpha_{H_{\hat{i}}} - \alpha_{L_{\hat{j}}} x\right)}$ as defined in (30).

By substituting (56 into (23) and using (25), the average BLER at user $H_{\hat{i}}$ in HCS method with TAS/SC or TAS/MRC is expressed as

$$\bar{\varepsilon}_{H_{\hat{i}}}^{HCS} \approx 1 + \chi_{H_{\hat{i}}}\sqrt{N_{H_{\hat{i}}}} \sum_{p=1}^{a_{H,I}} \sum_{\Delta_H=p} \Phi_H c_{H,I} \\ \times \int_{v_{H_{\hat{i}}}}^{\mu_{H_{\hat{i}}}} B_x^{\varphi_H} e^{-\frac{pm_H B_x}{\lambda_{SH}}} dx, \tag{57}$$

To derive the integral in (57), we carry out the change of variable by letting $t = B_x$ and (57) can be rewritten as

$$\bar{\varepsilon}_{H_{\hat{i}}}^{HCS} \approx 1 + \mathcal{A}_{H,1} \sum_{p=1}^{a_{H,I}} \sum_{\Delta_H=p} \Phi_H c_{H,I} \int_{B_{v_{H_{\hat{i}}}}}^{B_{\mu_{H_{\hat{i}}}}} \frac{t^{\varphi_H} e^{-\frac{pm_H t}{\lambda_{SH}}}}{\left(\frac{1}{\gamma_0\alpha_{L_{\hat{j}}}}+t\right)^2} dt. \tag{58}$$

By letting $u = \frac{1}{\gamma_0\alpha_{L_{\hat{j}}}} + t$ and using binomial expansion [53, Eq. (1.111)], (58) has the following form

$$\bar{\varepsilon}_{H_{\hat{i}}}^{HCS} \approx 1 + \mathcal{A}_{H,1}\widetilde{\sum}_{H,I} c_{H,I}\mathcal{A}_{H,2}\underbrace{\int_{\phi_{H_{\hat{i}}}}^{\kappa_{H_{\hat{i}}}} u^{\hat{\varphi}_H} e^{-\omega_H u} du}_{\mathcal{A}_{H,3}}. \tag{59}$$

We derive $\mathcal{A}_{H,3}$ in (59) with the aid of [53, Eqs. (3.351.4), (3.352.2), and (3.351.2)] and the final expression of $\bar{\varepsilon}_{H_{\hat{i}}}^{HCS}$ is achieved as in (30).

APPENDIX C
PROOF OF THEOREM 2

From (14) and (15), the CDF of $\gamma_{L_{\hat{j}}}^{x_{H_{\hat{i}}}}$ and $\gamma_{L_{\hat{j}}}^{x_{L_{\hat{j}}}}$ are, respectively, given by

$$F_{\gamma_{L_{\hat{j}}}^{x_{H_{\hat{i}}}}}(x) = \Pr\left\{\frac{\alpha_{H_{\hat{i}}}\gamma_0 g_{SL}}{\alpha_{L_{\hat{j}}}\gamma_0 g_{SL} + 1} < x\right\}$$
$$= F_{g_{SL}}(B_x), \tag{60}$$

and

$$F_{\gamma_{L_{\hat{j}}}^{x_{L_{\hat{j}}}}}(x) = \Pr\left\{\frac{\alpha_{L_{\hat{j}}}\gamma_0 g_{SL}}{\psi\alpha_{H_{\hat{i}}}\gamma_0 g_{SL} + 1} < x\right\}$$
$$= \begin{cases} F_{g_{SL}}\left(\frac{x}{\alpha_{L_{\hat{j}}}\gamma_0}\right), & \psi = 0 \\ F_{g_{SL}}\left(\hat{B}_x\right), & 0 < \psi \le 1 \end{cases}. \tag{61}$$

To derive $\bar{\varepsilon}_{L_{\hat{j}}}^{HCS}$ in (31), we need to resolve $\bar{\varepsilon}_{L_{\hat{j}}}^{x_{H_{\hat{i}}},HCS}$ and $\bar{\varepsilon}_{L_{\hat{j}}}^{x_{L_{\hat{j}}},HCS}$. For $\bar{\varepsilon}_{L_{\hat{j}}}^{x_{H_{\hat{i}}},HCS}$, from (26), (24), and (60), it can be expressed as

$$\bar{\varepsilon}_{L_{\hat{j}}}^{x_{H_{\hat{i}}},HCS} \approx 1 + \chi_{H_{\hat{i}}}\sqrt{N_{H_{\hat{i}}}}\sum_{p=1}^{a_{L,I}}\sum_{\Delta_L = p}\Phi_L c_{L,I}$$
$$\times \int_{v_{H_{\hat{i}}}}^{\mu_{H_{\hat{i}}}} B_x^{\varphi_L} e^{-\frac{pm_L B_x}{\lambda_{SL}}} dx. \tag{62}$$

After some algebraic manipulations similar to the proof of Theorem 1 in Appendix B, the final expression for $\bar{\varepsilon}_{L_{\hat{j}}}^{x_{H_{\hat{i}}},HCS}$ can be obtained as in (31).

For $\bar{\varepsilon}_{L_{\hat{j}}}^{x_{L_{\hat{j}}},HCS}$ in (31), we need to derive $\bar{\varepsilon}_{L_{\hat{j}},1}^{x_{L_{\hat{j}}},HCS}$ and $\bar{\varepsilon}_{L_{\hat{j}},2}^{x_{L_{\hat{j}}},HCS}$ to obtain its final expression. Specifically, with the aid of (24), (26), and (61) for $\psi = 0$, $\bar{\varepsilon}_{L_{\hat{j}},1}^{x_{L_{\hat{j}}},HCS}$ can be expressed as

$$\bar{\varepsilon}_{L_{\hat{j}},1}^{x_{L_{\hat{j}}},HCS} \approx 1 + \chi_{L_{\hat{j}}}\sqrt{N_{L_{\hat{j}}}}\sum_{p=1}^{a_{L,I}}\sum_{\Delta_L = p}\frac{\Phi_L c_{L,I}}{\left(\alpha_{L_{\hat{j}}}\gamma_0\right)^{\varphi_L}}$$
$$\times \int_{v_{L_{\hat{j}}}}^{\mu_{L_{\hat{j}}}} x^{\varphi_L} e^{-\hat{\omega}_L x} dx. \tag{63}$$

By using [53, Eq. (3.351.2)], the integral in (63) can be represented as

$$\int_{v_{L_{\hat{j}}}}^{\mu_{L_{\hat{j}}}} x^{\varphi_L} e^{-\hat{\omega}_L x} dx = \hat{\omega}_L^{-\varphi_L - 1}\Xi_{L,4}, \tag{64}$$

where $\Xi_{L,4}$ is defined in (31). By substituting (64) into (63), we obtain the final expression for $\bar{\varepsilon}_{L_{\hat{j}},1}^{x_{L_{\hat{j}}},HCS}$ as in (31).

By utilizing (24), (26), and (61) for the case $0 < \psi \le 1$, $\bar{\varepsilon}_{L_{\hat{j}},2}^{x_{L_{\hat{j}}},HCS}$ is given by

$$\bar{\varepsilon}_{L_{\hat{j}},2}^{x_{L_{\hat{j}}},HCS} \approx 1 + \chi_{L_{\hat{j}}}\sqrt{N_{L_{\hat{j}}}}\sum_{p=1}^{a_{L,I}}\sum_{\Delta_L = p}\Phi_L c_{L,I}$$
$$\times \int_{v_{L_{\hat{j}}}}^{\mu_{L_{\hat{j}}}} \hat{B}_x^{\varphi_L} e^{-\frac{pm_L \hat{B}_x}{\lambda_{SL}}} dx. \tag{65}$$

After some algebraic manipulations similar to the proof of Theorem 1 in Appendix B, the final expression for $\bar{\varepsilon}_{L_{\hat{j}},2}^{x_{L_{\hat{j}}},HCS}$ can be achieved as in (31).

REFERENCES

[1] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 124–130, Jun. 2018.

[2] P. Popovski, J. J. Nielsen, C. Stefanovic, E. de Carvalho, E. Strom, K. F. Trillingsgaard, A.-S. Bana, D. M. Kim, R. Kotaba, J. Park, and R. B. Sorensen, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Netw.*, vol. 32, no. 2, pp. 16–23, Mar. 2018.

[3] S. K. Sharma and X. Wang, "Toward massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 426–471, Firstquarter 2020.

[4] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint pilot and payload power allocation for massive-MIMO-enabled URLLC IIoT networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 816–830, May 2020.

[5] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Resource allocation for secure URLLC in mission-critical IoT scenarios," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5793–5807, Sep. 2020.

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

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

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

[9] M. Mousaei and B. Smida, "Optimizing pilot overhead for ultra-reliable short-packet transmission," in *IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017.

[10] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Nonorthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[11] L. Lei, L. You, Y. Yang, D. Yuan, S. Chatzinotas, and B. Ottersten, "Load coupling and energy optimization in multi-cell and multi-carrier NOMA networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 11 323–11 337, Nov. 2019.

[12] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, Thirdquarter 2018.

[13] A. C. Cirik, N. M. Balasubramanya, L. Lampe, G. Vos, and S. Bennett, "Toward the standardization of grant-free operation and the associated NOMA strategies in 3GPP," *IEEE Commun. Stand. Mag.*, vol. 3, no. 4, pp. 60–66, Dec. 2019.

[14] 3rd Generation Partnership Project (3GPP), "Study on downlink multiuser superposition transmission (MUST) for LTE," 3GPP, Tech. Rep. 36.859, Jan. 2016.

[15] 3rd Generation Partnership Project (3GPP), "Evolved universal terrestrial radio access (E-UTRA); Physical channels and modulation," 3GPP, Tech. Rep. TS 36.211, Apr. 2020.

[16] 3rd Generation Partnership Project (3GPP), "Study on non-orthogonal multiple access (NOMA) for NR," 3GPP, Tech. Rep. 38.812, Dec. 2018.

[17] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen, and Z. Zhong, "Short-packet downlink transmission with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4550–4564, Jul. 2018.

[18] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open J. Commun. Soc.*, no. 1, pp. 179–189, Jan. 2020.

[19] I. Budhiraja, N. Kumar, and S. Tyagi, "Cross-layer interference management scheme for D2D mobile users using NOMA," *IEEE Syst. J.*, pp. 1–12, Jun. 2020, Early Access.

[20] Z. Chang, L. Lei, H. Zhang, T. Ristaniemi, S. Chatzinotas, B. Ottersten, and Z. Han, "Secure and energy-efficient resource allocation for multiple-antenna NOMA with wireless power transfer," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 4, pp. 1059–1071, Dec. 2018.

[21] Q. Yu, C. Han, L. Bai, J. Wang, J. Choi, and X. Shen, "Multiuser selection criteria for MIMO-NOMA systems with different detectors," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1777–1791, Feb. 2020.

[22] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[23] Z. Wei, L. Zhao, J. Guo, D. W. K. Ng, and J. Yuan, "Multi-beam NOMA for hybrid mmWave systems," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1705–1719, Feb. 2019.

[24] M. Qiu, Y. Huang, and J. Yuan, "Downlink non-orthogonal multiple access without SIC for block fading channels: An algebraic rotation approach," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 3903–3918, Aug. 2019.

[25] S. Sanayei and A. Nosratinia, "Antenna selection in MIMO systems," *IEEE Commun. Mag.*, vol. 42, no. 10, pp. 68–73, Oct. 2004.

[26] Y. Yu, H. Chen, Y. Li, Z. Ding, and L. Zhuo, "Antenna selection in MIMO cognitive radio-inspired NOMA systems," *IEEE Commun. Lett.*, vol. 21, no. 12, pp. 2658–2661, Dec. 2017.

[27] Y. Yu, H. Chen, Y. Li, Z. Ding, L. Song, and B. Vucetic, "Antenna selection for MIMO nonorthogonal multiple access systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3158–3171, Apr. 2018.

[28] G. J. Sutton, J. Zeng, R. P. Liu, W. Ni, D. N. Nguyen, B. A. Jayawickrama, X. Huang, M. Abolhasan, Z. Zhang, E. Dutkiewicz, and T. Lv, "Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2488–2524, Feb. 2019.

[29] Y. Yu, H. Chen, Y. Li, Z. Ding, and B. Vucetic, "On the performance of non-orthogonal multiple access in short-packet communications," *IEEE Commun. Lett.*, vol. 22, no. 3, pp. 590–593, Mar. 2018.

[30] J. Zheng, Q. Zhang, and J. Qin, "Average block error rate of downlink NOMA short-packet communication systems in nakagami-m fading channels," *IEEE Commun. Lett.*, vol. 23, no. 10, pp. 1712–1716, Oct. 2019.

[31] X. Lai, Q. Zhang, and J. Qin, "Cooperative NOMA short-packet communications in flat Rayleigh fading channels," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 6182–6186, Jun. 2019.

[32] Y. Xu, C. Shen, D. Cai, and G. Zhu, "Latency constrained non-orthogonal packets scheduling with finite blocklength codes," *IEEE Trans. Veh. Technol.*, pp. 1–5, Jul. 2020, Early Access.

[33] Y. Xu, C. Shen, T. Chang, S. Lin, Y. Zhao, and G. Zhu, "Transmission energy minimization for heterogeneous low-latency NOMA downlink," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1054–1069, Feb. 2020.

[34] Z. Wang, T. Lv, Z. Lin, J. Zeng, and P. T. Mathiopoulos, "Outage performance of URLLC NOMA systems with wireless power transfer," *IEEE Wireless Commun. Lett.*, vol. 9, no. 3, pp. 380–384, Mar. 2020.

[35] C. Xiao, J. Zeng, W. Ni, X. Su, R. P. Liu, T. Lv, and J. Wang, "Downlink MIMO-NOMA for ultra-reliable low-latency communications," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 780–794, Apr. 2019.

[36] X. Huang and N. Yang, "On the block error performance of short-packet non-orthogonal multiple access systems," in *IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, May 2019.

[37] N. Yang, P. L. Yeoh, M. Elkashlan, R. Schober, and I. B. Collings, "Transmit antenna selection for security enhancement in MIMO wiretap channels," *IEEE Trans. Commun.*, vol. 61, no. 1, pp. 144–154, Jan. 2013.

[38] Z. Ding, L. Dai, and H. V. Poor, "MIMO-NOMA design for small packet transmission in the internet of things," *IEEE Access*, vol. 4, pp. 1393–1405, Apr. 2016.

[39] Z. Ding, H. Dai, and H. V. Poor, "Relay selection for cooperative NOMA," *IEEE Wireless Commun. Lett.*, vol. 5, no. 4, pp. 416–419, Jun. 2016.

[40] D.-D. Tran, D.-B. Ha, V. N. Vo, C. So-In, H. Tran, T. G. Nguyen, Z. A. Baig, and S. Sanguanpong, "Performance analysis of DF/AF cooperative MISO wireless sensor networks with NOMA and SWIPT over Nakagami-m fading," *IEEE Access*, vol. 6, pp. 56 142–56 161, Oct. 2018.

[41] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[42] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K.Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[43] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient joint user-RB association and power allocation for uplink hybrid NOMA-OMA," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5119–5131, Jun. 2019.

[44] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[45] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative nonorthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.

[46] J. Wang, B. Xia, K. Xiao, and Z. Chen, "Performance analysis and power allocation strategy for downlink NOMA systems in large-scale cellular networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3459–3464, Mar. 2020.

[47] N. T. Do, D. B. D. Costa, T. Q. Duong, and B. An, "A BNBF user selection scheme for NOMA-based cooperative relaying systems with SWIPT," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 664–667, Mar. 2017.

[48] D. C. González, D. B. da Costa, and J. C. S. S. Filho, "Distributed TAS/MRC and TAS/SC schemes for fixed-gain AF systems with multiantenna relay: Outage performance," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4380–4392, Jun. 2016.

[49] I. A. Mahady, E. Bedeer, S. Ikki, and H. Yanikomeroglu, "Sum-rate maximization of NOMA systems under imperfect successive interference cancellation," *IEEE Commun. Lett.*, vol. 23, no. 3, pp. 474–477, Mar. 2019.

[50] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of the incremental redundancy HARQ," *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 529–532, Oct. 2014.

[51] H. Ren, C. Pan, K. Wang, Y. Deng, M. Elkashlan, and A. Nallanathan, "Achievable data rate for URLLC-enabled UAV systems with 3-D channel model," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1587–1590, Dec. 2019.

[52] X. Zhang, X. Zhou, and M. R. McKay, "Enhancing secrecy with multi-antenna transmission in wireless ad hoc networks," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 11, pp. 1802–1814, Nov. 2013.

[53] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. Academic Press, Mar. 2007.

[54] C. Pan, H. Ren, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint blocklength and location optimization for URLLC-enabled UAV relay systems," *IEEE Commun. Lett.*, vol. 23, no. 3, pp. 498–501, Mar. 2019.

[55] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint power and blocklength optimization for URLLC in a factory automation scenario," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1786–1801, Mar. 2020.

Duc-Dung Tran (S'20) received the B.E. degree in electronics and telecommunications from Hue University of Sciences, Vietnam, in 2013, and the M.Sc. degree in computer sciences from Duy Tan University, Vietnam, in 2016. From 2014 to 2019, he was with Faculty of Electrical and Electronics Engineering, Duy Tan University. He is currently pursuing the Ph.D. degree with the Interdisciplinary Center for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg. His current research interests include 5G and beyond wireless networks, machine learning, URLLC, multiple access techniques, and wireless energy harvesting.

Shree Krishna Sharma (S'12–M'15–SM'18) is currently Research scientist at the Interdisciplinary Center for Security, Reliability and Trust (SnT), University of Luxembourg. Prior to this, he held research positions at the University of Western Ontario, Canada, and Ryerson University, Canada; and worked as Research Associate at the SnT after receiving his PhD degree in Wireless Communications from the University of Luxembourg in 2014. He has published more than 100 technical papers in scholarly journals, international conferences, and book chapters, and has over 3100 google scholar citations with an h-index of 26. His current research interests include 6G wireless, machine learning, Internet of Things, machine-type communications, edge computing and optimization of distributed communications, computing and caching resources.

He is a Senior Member of IEEE and is the recipient of several prestigious awards including "FNR Award for Outstanding PhD Thesis 2015" from FNR, Luxembourg, "Best Paper Award" in CROWNCOM 2015 conference, "2018 EURASIP JWCN Best Paper Award" and the co-recipient of "FNR Award for Outstanding Scientific Publication 2019". He has been serving as a Reviewer for several international journals and conferences; as a TPC member for a number of international conferences including IEEE ICC, IEEE GLOBECOM, IEEE PIMRC, IEEE VTC and IEEE ISWCS; and is in the Editorial board of IEEE Access and IEEE TCCN. He co-organized a special session in IEEE PIMRC 2017, a workshop in IEEE SECON 2019, was a Track co-chair for IEEE VTC-fall 2018 conference, and published an IET book on "Satellite Communications in the 5G Era" as a lead editor.

Symeon Chatzinotas is currently Full Professor/Chief Scientist I and Co-Head of the SIGCOM Research Group at SnT, University of Luxembourg. In the past, he has been a Visiting Professor at the University of Parma, Italy and he was involved in numerous Research and Development projects for the National Center for Scientific Research Demokritos, the Center of Research and Technology Hellas and the Center of Communication Systems Research, University of Surrey. He received the M.Eng. degree in telecommunications from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2003, and the M.Sc. and Ph.D. degrees in electronic engineering from the University of Surrey, Surrey, U.K., in 2006 and 2009, respectively. He was a co-recipient of the 2014 IEEE Distinguished Contributions to Satellite Communications Award, the CROWNCOM 2015 Best Paper Award and the 2018 EURASIC JWCN Best Paper Award. He has (co-)authored more than 400 technical papers in refereed international journals, conferences and scientific books. He is currently in the editorial board of the IEEE Open Journal of Vehicular Technology and the International Journal of Satellite Communications and Networking.

Isaac Woungang (Senior Member, IEEE) received the Ph.D. degree in mathematics from the University of the South, Toulon-Var, France, in 1994. From 1999 to 2002, he worked as Software Engineer at Nortel Networks Corporation, Ottawa, ON, Canada. Since 2002, he has been with Ryerson University, Toronto, ON, Canada, where he is currently a Professor of computer science. His current research interests include radio resource management in next generation wireless networks, big data, the Internet of Things (IoT), and cloud computing. He has published eight books and over 90 refereed technical papers in scholarly international journals and proceedings of international conferences. He has served as the Chair of the Computer Chapter, IEEE Toronto Section, from 2012 to 2018. He has Guest Edited several special issues with various reputed journals, such as Computer Communications (Elsevier) and Telecommunication Systems (Springer).

Björn Ottersten (S'87–M'89–SM'99–F'04) was born in Stockholm, Sweden, in 1961. He received the M.S. degree in electrical engineering and applied physics from Linköping University, Linköping, Sweden, in 1986, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1990. He has held research positions with the Department of Electrical Engineering, Linköping University, the Information Systems Laboratory, Stanford University, the Katholieke Universiteit Leuven, Leuven, Belgium, and the University of Luxembourg, Luxembourg. From 1996 to 1997, he was the Director of Research with ArrayComm, Inc., a start-up in San Jose, CA, USA, based on his patented technology. In 1991, he was appointed Professor of signal processing with the Royal Institute of Technology (KTH), Stockholm, Sweden. Dr. Ottersten has been Head of the Department for Signals, Sensors, and Systems, KTH, and Dean of the School of Electrical Engineering, KTH. He is currently the Director for the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg.

He is a recipient of the IEEE Signal Processing Society Technical Achievement Award and the European Research Council advanced research grant twice. He has co-authored journal papers that received the IEEE Signal Processing Society Best Paper Award in 1993, 2001, 2006, 2013, and 2019, and 8 IEEE conference papers best paper awards. He has been a board member of IEEE Signal Processing Society, the Swedish Research Council and currently serves of the boards of EURASIP and the Swedish Foundation for Strategic Research. He has served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the Editorial Board of the IEEE Signal Processing Magazine. He is currently a member of the editorial boards of IEEE Open Journal of Signal Processing, EURASIP Signal Processing Journal, EURASIP Journal of Advances Signal Processing and Foundations and Trends of Signal Processing. He is a fellow of EURASIP.