OTFS-NOMA: An Efficient Approach for Exploiting Heterogenous User Mobility Profiles

DOI:
10.1109/TCOMM.2019.2932934

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your article. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

AQ:1 = Please supply index terms/keywords for your paper. To download the IEEE Taxonomy, go to http://www.ieee.org/documents/taxonomy_v101.pdf.

AQ:2 = Author: Please confirm or add details for any funding or financial support for the research of this article.

AQ:3 = Please provide the postal code for The University of Manchester, Friedrich-Alexander-University Erlangen-Nurnberg, and Southwest Jiaotong University.

AQ:4 = Note that if you require corrections/changes to tables or figures, you must supply the revised files, as these items are not edited for you.

AQ:5 = Please provide the department name for Ref. [37].

AQ:6 = Current affiliation in the biography of Zhiguo Ding does not match the First Footnote. Please check and correct where needed.

OTFS-NOMA: An Efficient Approach for Exploiting Heterogenous User Mobility Profiles

Zhiguo Ding, *Senior Member, IEEE*, Robert Schober, *Fellow, IEEE*, Pingzhi Fan, *Fellow, IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*— This paper considers a challenging communication scenario, in which users have heterogenous mobility profiles, e.g., some users are moving at high speeds and some users are static. A new non-orthogonal multiple-access (NOMA) transmission protocol that incorporates orthogonal time frequency space (OTFS) modulation is proposed. Thereby, users with different mobility profiles are grouped together for the implementation of NOMA. The proposed OTFS-NOMA protocol is shown to be applicable to both uplink and downlink transmission, where sophisticated transmit and receive strategies are developed to remove inter-symbol interference and harvest both multi-path and multi-user diversity. Analytical results demonstrate that both the high-mobility and the low-mobility users benefit from the application of OTFS-NOMA. In particular, the use of NOMA allows the spreading of the high-mobility users' signals over a large amount of time-frequency resources, which enhances the OTFS resolution and improves the detection reliability. In addition, OTFS-NOMA ensures that low-mobility users have access to bandwidth resources which in conventional OTFS-orthogonal multiple access (OTFS-OMA) would be solely occupied by the high-mobility users. Thus, OTFS-NOMA improves the spectral efficiency and reduces latency.**

AQ:1 ***Index Terms*— XXXXX.**

I. Introduction

NON-ORTHOGONAL multiple access (NOMA) has been recognized as a paradigm shift for the design of multiple access techniques for the next generation of wireless networks [1]–[4]. Many existing works on NOMA have focused on scenarios with low-mobility users, where users with different channel conditions or quality of service (QoS) requirements are grouped together for the implementation of NOMA. For example, in power-domain NOMA, a base station serves two users simultaneously [5], [6]. In particular, the base station first orders the users according to their channel conditions, where the 'weak user' which has a poorer connection to the base station is generally allocated more transmission power and the other user, referred to as the 'strong user', is allocated less power. As such, the two users can be served in the same time-frequency resource, which improves the spectral efficiency compared to orthogonal multiple access (OMA). In the case that users have similar channel conditions, grouping users with different QoS requirements can facilitate the implementation of NOMA and effectively exploit the potential of NOMA [7]–[9]. Various existing studies have shown that the NOMA principle can be applied to different communication networks, such as millimeter-wave networks [10], [11], massive multiple-input multiple-output (MIMO) systems [12], [13], hybrid multiple access systems [14], [15], visible light communication networks [16], [17], and mobile edge computing [18]. We also note that various standardization efforts have been made to facilitate the implementation of NOMA in practical systems. For example, a study for the application of NOMA for downlink transmission, termed multi-user superposition transmission (MUST), was carried out for the 3rd Generation Partnership Project (3GPP) Release 14, where 15 different forms of MUST were proposed and compared [19]. After this study was completed, MUST was formally included in 3GPP Release 15 which is also referred to as Evolved Universal Terrestrial Radio Access (E-UTRA) [20]. A study for the application of NOMA for uplink transmission has been recently carried out for 3GPP Release 16, where more than 20 different forms of NOMA have been proposed by various companies [21].

This paper considers the application of NOMA to a challenging communication scenario, where users have heterogeneous mobility profiles. Different from the existing works in [22], [23], the use of orthogonal time frequency space (OTFS) modulation is considered in this paper because of its superior performance in scenarios with doubly-dispersive channels [24]–[26]. Recall that the key idea of OTFS is to use the delay-Doppler plane, where the users' signals are

Manuscript received April 4, 2019; revised June 5, 2019; accepted
AQ:2 July 14, 2019. The work of Z. Ding was supported by the UK Engineering and Physical Sciences Research Council under grant number EP/P009719/2 and by H2020-MSCA-RISE-2015 under grant number 690750. The work of P. Fan was supported by the National Natural Science Foundation of China under grant number 61731017, and the 111 Project (No.111-2-14). The work of H. V. Poor was supported by the U.S. National Science Foundation under Grants CCF-093970 and CCF-1513915. The associate editor coordinating the review of this article and approving it for publication was V. Raghavan. *(Corresponding author: Zhiguo Ding.)*

Z. Ding is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA, and also with the School of Electrical
AQ:3 and Electronic Engineering, The University of Manchester, Manchester, U.K. (e-mail: zhiguo.ding@manchester.ac.uk).

R. Schober is with the Institute for Digital Communications, Friedrich-Alexander-University Erlangen-Nurnberg (FAU), Erlangen, Germany (e-mail: robert.schober@fau.de).

P. Fan is with the Institute of Mobile Communications, Southwest Jiaotong University, Chengdu, China (e-mail: pingzhifan@foxmail.com).

H. V. Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Digital Object Identifier 10.1109/TCOMM.2019.2932934

orthogonally placed. Compared to conventional modulation schemes, such as orthogonal frequency-division multiplexing (OFDM), OTFS offers the benefit that the time-invariant channel gains in the delay-Doppler plane can be utilized, which simplifies channel estimation and signal detection in high-mobility scenarios. The impact of pulse-shaping waveforms on the performance of OTFS was studied in [27], and the design of interference cancellation and iterative detection for OTFS was investigated in [28]. The diversity gain achieved by OTFS was studied in [29], and the application of OTFS to multiple access was proposed in [30]. In [31] and [32], the concept of OTFS was combined with MIMO, which revealed that the use of spatial degrees of freedom can further enhance the performance of OTFS.

This paper considers the application of OTFS to NOMA communication networks, where the coexistence of NOMA and OTFS is investigated. In particular, this paper makes the following contributions:

1) A spectrally efficient OTFS-NOMA transmission protocol is proposed by grouping users with different mobility profiles for the implementation of NOMA. On the one hand, users with high mobility are served in the delay-Doppler plane, and their signals are modulated by OTFS. On the other hand, users with low mobility are served in the time-frequency plane, and their signals are modulated in a manner similar to conventional OFDM.
2) The proposed new OTFS-NOMA protocol is applied to both uplink and downlink transmission, where different rate and power allocation policies are used to suppress multiple access interference. In addition, sophisticated equalization techniques, such as the frequency-domain zero-forcing linear equalizer (FD-LE) and the decision feedback equalizer (FD-DFE), are employed to remove the inter-symbol interference in the delay-Doppler plane. The impact of the developed equalization techniques on OTFS-NOMA is analyzed by using the outage probability as the performance criterion. Strategies to harvest multi-path diversity and multi-user diversity are also introduced, which can further improve the outage performance of OTFS-NOMA transmission.
3) The developed analytical results demonstrate that both the high-mobility and the low-mobility users benefit from the proposed OTFS-NOMA scheme. The use of NOMA allows the high-mobility users' signals to be spread over a large amount of time-frequency resources without degrading the spectral efficiency. As a result, the OTFS resolution, which determines whether the users' channels can be accurately located in the delay-Doppler plane, is enhanced significantly, and therefore, the reliability of detecting the high-mobility users' signals is improved. We note that, in OTFS-OMA, enhancing the OTFS resolution implies that a large amount of time and frequency resources are solely occupied by the high-mobility users, which reduces the overall spectral efficiency since the high-mobility users' channel conditions are typically weaker than those of the low-mobility users. In contrast, the use of OTFS-NOMA ensures that the low-mobility users can access the bandwidth resources which would be solely occupied by the high-mobility users in the OMA mode. Hence, OTFS-NOMA improves spectral efficiency and reduces latency, as with OTFS-OMA the low-mobility users may have to wait for a long time before the scarce bandwidth resources occupied by the high-mobility users become available. In addition, we note that for the low-mobility users, using OFDM yields the same reception reliability as using OTFS, as pointed out in [33]. Therefore, the proposed OTFS-NOMA scheme, which serves the low-mobility users in the time-frequency plane and modulates the low-mobility users' signals in a manner similar to OFDM, offers the same reception reliability as OTFS-OMA, which serves the low-mobility users in the delay-Doppler plane and modulates the low-mobility users' signals by OTFS. However, OTFS-NOMA has the benefit of reduced system complexity because the use of the complicated OTFS transforms is avoided.

II. Foundations of OTFS-NOMA

A. Time-Frequency Plane and Delay-Doppler Plane

The key idea of OTFS-NOMA is to efficiently use both the time-frequency plane and the delay-Doppler plane. A discrete time-frequency plane is obtained by sampling at intervals of T s and Δf Hz as follows:

$$\Lambda_{\text{TF}}=\{(nT, m\Delta f), n=0,\cdots,N-1, m=0,\cdots,M-1\}, \tag{1}$$

and the corresponding discrete delay-Doppler plane is given by

$$\Lambda_{\text{DD}}=\left\{\left(\frac{k}{NT},\frac{l}{M\Delta f}\right), k=0,\cdots,N-1, l=0,\cdots,M-1\right\}, \tag{2}$$

where N and M denote the total number of time intervals and the total number of frequency subchannels, respectively. The choices for T and Δf are determined by the channel characteristics, as will be explained in the following subsection.

B. Channel Model

This paper considers a multi-user communication network in which one base station communicates with $(K+1)$ users, denoted by U_i, $0 \leq i \leq K$. Denote U_i's channel response in the delay-Doppler plane by $h_i(\tau,\nu)$, where τ denotes the delay and ν denotes the Doppler shift. OTFS uses the sparsity feature of a wireless channel in the delay-Doppler plane, i.e., there are a small number of propagation paths between a transmitter and a receiver [24], [25], [28], which means that $h_i(\tau,\nu)$ can be expressed as follows:

$$h_i(\tau,\nu)=\sum_{p=0}^{P_i} h_{i,p}\delta(\tau-\tau_{i,p})\delta(\nu-\nu_{i,p}), \tag{3}$$

where (P_i+1) denotes the number of propagation paths, and $h_{i,p}$, $\tau_{i,p}$, and $\nu_{i,p}$ denote the complex Gaussian channel gain,[1] the delay, and the Doppler shift associated with the

[1]The Gaussian assumption has been commonly used in the OTFS literature [26]–[29] since each channel gain (or each tap of the delay-Doppler impulse response) represents a cluster of reflectors with specific delay and Doppler characteristics.

p-th propagation path, respectively. We assume that the $h_{i,p}$, $0 \leq p \leq P_i$, are independent and identically distributed (i.i.d.) random variables,[2] i.e., $h_{i,p} \sim CN\left(0, \frac{1}{P_i+1}\right)$, which means $\sum_{p=0}^{P_i} \mathcal{E}\{|h_{i,p}|^2\} = 1$, where $\mathcal{E}\{\cdot\}$ denotes the expectation operation. The discrete delay and Doppler tap indices for the p-th path of $h_i(\tau, \nu)$, denoted by $l_{\tau_{i,p}}$ and $k_{\nu_{i,p}}$, respectively, are given by [28]

$$\tau_{i,p} = \frac{l_{\tau_{i,p}} + \tilde{l}_{\tau_{i,p}}}{M\Delta f}, \quad \nu_{i,p} = \frac{k_{\nu_{i,p}} + \tilde{k}_{\nu_{i,p}}}{NT}, \tag{4}$$

where $\tilde{l}_{\tau_{i,p}}$ and $\tilde{k}_{\nu_{i,p}}$ denote the fractional delay and the fractional Doppler shift, respectively.

The construction of Λ_{TF} and Λ_{DD} needs to ensure that T is not smaller than the maximal delay spread, and Δf is not smaller than the largest Doppler shift, i.e., $T \geq \max\{\tau_{i,p}, 0 \leq p \leq P_i, 0 \leq i \leq K\}$ and $\Delta f \geq \max\{\nu_{i,p}, 0 \leq p \leq P_i, 0 \leq i \leq K\}$. In addition, the choices of N and M affect the OTFS resolution, which determines whether $h_i(\tau, \nu)$ can be accurately located in the discrete delay-Doppler plane. In particular, M and N need to be sufficiently large to approximately achieve ideal OTFS resolution, which ensures that $\tilde{l}_{\tau_{i,p}} = \tilde{k}_{\nu_{i,p}} = 0$, such that the interference caused by fractional delay and Doppler shift is effectively suppressed [24].

C. General Principle of OTFS-NOMA

To facilitate the illustration of the general principle of OTFS-NOMA, we first briefly describe OTFS-OMA, the benchmark scheme used in this paper. In OTFS-OMA, there is no spectrum sharing between the high-mobility users and the low-mobility users, i.e., if OTFS is used to serve the high-mobility users, the NT time intervals and the $M\Delta f$ frequency subchannels are occupied by the high-mobility users and the low-mobility users cannot be served in these resource blocks. The general principle of the proposed OTFS-NOMA scheme is to exploit both the delay-Doppler plane and the time-frequency plane, where users with heterogenous mobility profiles are grouped together and served simultaneously. On the one hand, for the users with high mobility, their signals are placed in the delay-Doppler plane, which means that the time-invariant channel gains in the delay-Doppler plane can be exploited. It is worth pointing out that in order to ensure that the channels can be located in the delay-Doppler plane, both N and M need to be large, which is a disadvantage of OTFS-OMA, since a significant number of frequency channels (e.g., $M\Delta f$) are occupied for a long time (e.g., NT) by the high-mobility users whose channel conditions can be quite weak. The use of OTFS-NOMA facilitates spectrum sharing and hence ensures that the high-mobility users' signals can be spread over a large amount of time-frequency resources without degrading the spectral efficiency.

On the other hand, for the users with low mobility, their signals are placed in the time-frequency plane. The interference between the users with different mobility profiles is managed by using the principle of NOMA. As a result, compared to OTFS-OMA, OTFS-NOMA improves the overall spectral efficiency since it encourages spectrum sharing among users with different mobility profiles and avoids that the bandwidth resources are solely occupied by the high-mobility users which might have weak channel conditions. In addition, the complexity of detecting the low-mobility users' signals is reduced, compared to OTFS-OMA which serves all users in the delay-Doppler plane.

In this paper, we assume that, among $(K+1)$ users, U_0 is a user with high mobility, and the remaining K users, U_i for $1 \leq i \leq K$, are low-mobility users, which are referred to as 'NOMA' users.[3] For OTFS-OMA, we assume that U_0 solely occupies all NM resource blocks in Λ_{DD}. In OTFS-NOMA, U_i, for $1 \leq i \leq K$, are opportunistic NOMA users and their signals are placed in Λ_{TF}. The design of downlink OTFS-NOMA transmission will be discussed in detail in Sections III, IV, and V. The application of OTFS-NOMA for uplink transmission will be considered in Section VI only briefly, due to space limitations.

III. Downlink OFTS-NOMA - System Model

In this section, the OTFS-NOMA downlink transmission protocol is described. In particular, assume that the base station sends NM symbols to U_0, denoted by $x_0[k,l]$, $k \in \{0, \cdots, N-1\}$, $l \in \{0, \cdots, M-1\}$. By using the inverse symplectic finite Fourier transform (ISFFT), the high-mobility user's symbols placed in the delay-Doppler plane are converted to NM symbols in the time-frequency plane as follows [24]:

$$X_0[n,m] = \frac{1}{NM} \sum_{k=0}^{N-1} \sum_{l=0}^{M-1} x_0[k,l] e^{j2\pi\left(\frac{kn}{N} - \frac{ml}{M}\right)}, \tag{5}$$

where $n \in \{0, \cdots, N-1\}$ and $m \in \{0, \cdots, M-1\}$. We note that the NM time-frequency signals can be viewed as N OFDM symbols containing M signals each. We assume that a rectangular window is applied to the transmitted and received signals.

The NOMA users' signals are placed directly in the time-frequency plane, and are superimposed with the high-mobility user's signals, $X_0[n,m]$. With NM orthogonal resource blocks available in the time-frequency plane, there are different ways for the K users to share the resource blocks. For illustration purposes, we assume that M users are selected from the K opportunistic NOMA users,[4] where each NOMA

[2] In order to simplify the performance analysis, we assume that the users' channels are i.i.d. In practice, it is likely that the high-mobility users' channel conditions are worse than the low-mobility users' channel conditions. This channel difference is beneficial for the implementation of NOMA, and hence can further increase the performance gain of OTFS-NOMA over OTFS-OMA.

[3] We note that the principle of OTFS-NOMA can be extended to the case where multiple high-mobility users are served in the delay-Doppler plane. In this case, the NM signals in the delay-Doppler plane belong to different high-mobility users and OTFS is used as a type of multiple access technique [24], [30]. For downlink transmission, this change has no impact on the proposed detection schemes and the analytical results developed in this paper. For uplink transmission, the results developed in this paper are applicable to the case with multiple high-mobility users if the adaptive-rate transmission scheme proposed in Section VI is employed.

[4] The same M users can be scheduled as long as the users' channels do not change in the delay-Doppler plane. Otherwise, a new set of M users may be selected from the K opportunistic users. We also note that the number of the opportunistic users is assumed to be larger than the number of the frequency subchannels ($K \geq M$), which can be justified by a spectrum crunch scenario, i.e., there are not sufficient bandwidth resources available to support a large number of mobile devices.

user is to occupy one frequency subchannel and receive N information bearing symbols, denoted by $x_i(n)$, for $1 \leq i \leq M$ and $0 \leq n \leq N-1$. The criterion employed for user scheduling and its impact on the performance of OTFS-NOMA will be discussed in Section V. Denote the time-frequency signals to be sent to U_i by $X_i[n,m]$, $1 \leq i \leq M$. The following mapping scheme is used in this paper[5]:

$$X_i[n,m] = \begin{cases} x_i(n) & \text{if} \quad m = i-1 \\ 0 & \text{otherwise}, \end{cases} \tag{6}$$

for $1 \leq i \leq M$ and $0 \leq n \leq N-1$.

The base station superimposes U_0's time-frequency signals with the NOMA users' signals as follows:

$$X[n,m] = \frac{\gamma_0}{NM} \sum_{k=0}^{N-1} \sum_{l=0}^{M-1} x_0[k,l] e^{j2\pi\left(\frac{kn}{N} - \frac{ml}{M}\right)} + \sum_{i=1}^{M} \gamma_i X_i[n,m], \tag{7}$$

where γ_i denotes the NOMA power allocation coefficient of user i, and $\sum_{i=0}^{M} \gamma_i^2 = 1$.

The transmitted signal at the base station is obtained by applying the Heisenberg transform to $X[n,m]$. By assuming perfect orthogonality between the transmit and receive pulses, the received signal at U_i in the time-frequency plane can be modelled as follows [24], [25], [28]:

$$Y_i[n,m] = H_i[n,m] X[n,m] + W_i[n,m], \tag{8}$$

where $W_i(n,m)$ is the white Gaussian noise in the time-frequency plane, and $H_i(n,m) = \int\int h_i(\tau,\nu) e^{j2\pi\nu nT} e^{-j2\pi(\nu + m\Delta f)\tau} d\tau d\nu$.

IV. Downlink OTFS-NOMA - Detecting the High-Mobility User's Signals

For the proposed downlink OTFS-NOMA scheme, U_0 directly detects its signals in the delay-Doppler plane by treating the NOMA users' signals as noise. In particular, in order to detect U_0's signals, the symplectic finite Fourier transform (SFFT) is applied to $Y_0[n,m]$ to obtain the delay-Doppler estimates as follows:

$$\begin{aligned} y_0[k,l] &= \frac{1}{NM} \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} Y_0[n,m] e^{-j2\pi\left(\frac{nk}{N} - \frac{ml}{M}\right)} \\ &= \frac{1}{NM} \sum_{q=0}^{M} \gamma_q \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} x_q[n,m] h_{w,0}\left(\frac{k-n}{NT}, \frac{l-m}{M\Delta f}\right) \\ &\quad + z_0[k,l], \end{aligned} \tag{9}$$

where q denotes the user index, $z_0[k,l]$ is complex Gaussian noise, $x_q[k,l]$, $1 \leq q \leq M$, denotes the delay-Doppler representation of $X_q[n,m]$ and can be obtained by applying the SFFT to $X_q[n,m]$, the channel $h_{w,0}(\nu',\tau')$ is given by

$$h_{w,0}(\nu',\tau') = \int\int h_i(\tau,\nu) w(\nu'-\nu, \tau'-\tau) e^{-j2\pi\nu\tau} d\tau d\nu, \tag{10}$$

and $w(\nu,\tau) = \sum_{c=0}^{N-1} \sum_{d=0}^{M-1} e^{-j2\pi(\nu cT - \tau d\Delta f)}$. To simplify the analysis, the power of the complex-Gaussian distributed noise is assumed to be normalized, i.e., $z_i[k,l] \sim CN(0,1)$, where $CN(a,b)$ denotes a complex Gaussian distributed random variable with mean a and variance b.

By applying the channel model in (3), the relationship between the transmitted signals and the observations in the delay-Doppler plane can be expressed as follows [24], [25], [28]:

$$y_0[k,l] = \sum_{q=0}^{M} \gamma_q \sum_{p=0}^{P_0} h_{0,p} x_q[(k - k_{\nu_{0,p}})_N, (l - l_{\tau_{0,p}})_M] + z_0[k,l], \tag{11}$$

where $(\cdot)_N$ denotes the modulo N operator. As in [29]–[31], we assume that N and M are sufficiently large to ensure that both $\tilde{k}_{\nu_{0,p}}$ and $\tilde{l}_{\tau_{0,p}}$ are zero, i.e., there is no interference caused by fractional delay or fractional Doppler shift. We note that for OTFS-OMA, increasing N and M can significantly reduce spectral efficiency, whereas the use of large N and M becomes possible for OTFS-NOMA because of the spectrum sharing of users with different mobility profiles.

Define $\mathbf{y}_{0,k} = \left[y_0[k,0] \cdots y_0[k,M-1]\right]^T$ and $\mathbf{y}_0 = \left[\mathbf{y}_{0,0}^T \cdots \mathbf{y}_{0,N-1}^T\right]^T$. Similarly, the signal vector $\mathbf{x}_i$ and the noise vector $\mathbf{z}_0$ are constructed from $x_i[k,l]$ and $z_0[k,l]$, respectively. Based on (11), the system model can be expressed in matrix form as follows:

$$\mathbf{y}_0 = \gamma_0 \mathbf{H}_0 \mathbf{x}_0 + \underbrace{\sum_{q=1}^{M} \gamma_q \mathbf{H}_0 \mathbf{x}_q + \mathbf{z}_0}_{\text{Interference and noise terms}}, \tag{12}$$

where $\mathbf{H}_0$ is a block-circulant matrix and defined as follows:

$$\mathbf{H}_0 = \begin{bmatrix} \mathbf{A}_{0,0} & \mathbf{A}_{0,N-1} & \cdots & \mathbf{A}_{0,2} & \mathbf{A}_{0,1} \\ \mathbf{A}_{0,1} & \mathbf{A}_{0,0} & \ddots & \mathbf{A}_{0,3} & \mathbf{A}_{0,2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{A}_{0,N-2} & \mathbf{A}_{0,N-3} & \ddots & \mathbf{A}_{0,0} & \mathbf{A}_{0,N-1} \\ \mathbf{A}_{0,N-1} & \mathbf{A}_{0,N-2} & \ddots & \mathbf{A}_{0,1} & \mathbf{A}_{0,0} \end{bmatrix}, \tag{13}$$

and each submatrix $\mathbf{A}_{0,n}$ is an $M \times M$ circulant matrix whose structure is determined by (11).

Example: Consider a special case with $N = 4$ and $M = 3$, and U_0's channel is given by

$$h_0(\tau,\nu) = h_{0,0}\delta(\tau)\delta(\nu) + h_{0,1}\delta\left(\tau - \frac{1}{M\Delta f}\right)\delta\left(\nu - \frac{3}{NT}\right), \tag{14}$$

which means $k_0 = 0$, $k_1 = 3$, $l_0 = 0$, $l_1 = 1$. Therefore, the block-circulant matrix is given by

$$\mathbf{H}_0 = \begin{bmatrix} \mathbf{A}_{0,0} & \mathbf{A}_{0,3} & \mathbf{A}_{0,2} & \mathbf{A}_{0,1} \\ \mathbf{A}_{0,1} & \mathbf{A}_{0,0} & \mathbf{A}_{0,3} & \mathbf{A}_{0,2} \\ \mathbf{A}_{0,2} & \mathbf{A}_{0,1} & \mathbf{A}_{0,0} & \mathbf{A}_{0,3} \\ \mathbf{A}_{0,3} & \mathbf{A}_{0,2} & \mathbf{A}_{0,1} & \mathbf{A}_{0,0} \end{bmatrix}, \tag{15}$$

[5] We note that mapping schemes different from (6) can also be used. For example, if N users are scheduled and each user is to occupy one time slot and receives an OFDM-like symbol containing M signals, we can set $X_i[n,m] = x_i(m)$, for $n = i-1$.

where $\mathbf{A}_{0,0} = h_{0,0}\mathbf{I}_3$, $\mathbf{A}_{0,1} = \mathbf{A}_{0,2} = \mathbf{0}_{3\times 3}$ and $\mathbf{A}_{0,3} = \begin{bmatrix} 0 & 0 & h_{0,1} \\ h_{0,1} & 0 & 0 \\ 0 & h_{0,1} & 0 \end{bmatrix}$.

Remark 1: It is well known that an $n\times n$ circulant matrix can be diagonalized by the $n\times n$ discrete Fourier transform (DFT) and inverse DFT matrices, denoted by $\mathbf{F}_n$ and $\mathbf{F}_n^{-1}$, respectively, i.e., the columns of the DFT matrix are the eigenvectors of the circulant matrix. We note that directly applying the DFT factorization to $\mathbf{H}_0$ is not possible, since $\mathbf{H}_0$ is not a circulant matrix, but a block circulant matrix.

Because of the structure of $\mathbf{H}_0$, inter-symbol interference still exists in the considered OTFS-NOMA system, and equalization is needed. We consider two equalization approaches, FD-LE and FD-DFE, which were both originally developed for single-carrier transmission with cyclic prefix [34], [35].

A. Design and Performance of FD-LE

The proposed FD-LE consists of two steps. Let $\otimes$ denote the Kronecker product. The first step is to multiply the observation vector $\mathbf{y}_0$ by $\mathbf{F}_N \otimes \mathbf{F}_M^H$, which leads to the result in the following proposition.

Proposition 1: By applying the detection matrix $\mathbf{F}_N \otimes \mathbf{F}_M^H$ to observation vector $\mathbf{y}_0$, the received signals for OTFS-NOMA downlink transmission can be written as follows:

$$\tilde{\mathbf{y}}_0 = \mathbf{D}_0(\mathbf{F}_N \otimes \mathbf{F}_M^H)\left(\gamma_0\mathbf{x}_0 + \sum_{q=1}^{M}\gamma_q\mathbf{x}_q\right) + \tilde{\mathbf{z}}_0, \quad (16)$$

where $\tilde{\mathbf{y}}_0 = (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{y}_0$, $\tilde{\mathbf{z}}_0 = (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{z}_0$, $\mathbf{D}_0$ is a diagonal matrix whose $(kM+l+1)$-th main diagonal element is given by

$$D_0^{k,l} = \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} a_{0,n}^{m,1} e^{j2\pi\frac{lm}{M}} e^{-j2\pi\frac{kn}{N}}, \quad (17)$$

for $0 \le k \le N-1$, $0 \le l \le M-1$, and $a_{0,n}^{m,1}$ is the element located in the $(nM+m+1)$-th row and the first column of $\mathbf{H}_0$.

Proof: Please refer to Appendix A. □

With the simplified signal model shown in (16), the second step of FD-LE is to apply $\left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)^{-1}\mathbf{D}_0^{-1}$ to $\tilde{\mathbf{y}}_0$. Thus, U_0's received signal is given by

$$\check{\mathbf{y}}_0 = \gamma_0\mathbf{x}_0 + \underbrace{\sum_{q=1}^{M}\gamma_q\mathbf{x}_q + \left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)^{-1}\mathbf{D}_0^{-1}\tilde{\mathbf{z}}_0}_{\text{Interference and noise terms}}, \quad (18)$$

where $\check{\mathbf{y}}_0 = \left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)^{-1}\mathbf{D}_0^{-1}\tilde{\mathbf{y}}_0$. To simplify the analysis, we assume that the powers of all users' information-bearing signals are identical, which means that the transmit signal-to-noise ratio (SNR) can be defined as $\rho = \mathcal{E}\{|x_0[k,l]|^2\} = \mathcal{E}\{|x_i(n)|^2\}$, since the noise power is assumed to be normalized.[6] The following lemma provides the signal-to-interference-plus-noise ratio (SINR) achieved by FD-LE.

[6]Following steps in the proof for Proposition 1 to show the statistical property of $\tilde{\mathbf{z}}_0$ in (66), we can also show that $W_i[n,m] \sim CN(0,1)$ if $z_i[k,l] \sim CN(0,1)$.

Lemma 1: Assume that $\gamma_i = \gamma_1$, for $1 \le i \le N$. By using FD-LE, the SINRs for detecting all $x_0[k,l]$, $0 \le k \le N-1$ and $0 \le l \le M-1$, are identical and given by

$$SINR_{0,kl}^{LE} = \frac{\rho\gamma_0^2}{\rho\gamma_1^2 + \frac{1}{NM}\sum_{\tilde{k}=0}^{N-1}\sum_{\tilde{l}=0}^{M-1}|D_0^{\tilde{k},\tilde{l}}|^{-2}}. \quad (19)$$

Proof: Please refer to Appendix B. □

Remark 2: The proof of Lemma 1 shows that $\sum_{i=0}^{M}\gamma_i^2 = 1$ can be simplified as $\gamma_0^2 + \gamma_i^2 = 1$ for $1 \le i \le M$, which is the motivation for assuming $\gamma_i = \gamma_1$. Following steps similar to those in the proofs for Proposition 1 and Lemma 1, one can show that directly applying $\mathbf{H}_0^{-1}$ to the observation vector yields the same SINR. However, the proposed FD-LE scheme can be implemented more efficiently since $\left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)^{-1} = \mathbf{F}_N^H \otimes \mathbf{F}_M$ and $\mathbf{D}_0$ is a diagonal matrix. Hence, the inversion of a full $NM \times NM$ matrix is avoided.

In this paper, the outage probability and the outage rate are used as performance criteria, since the outage probability can provide a tight bound on the probability of erroneous detection and is general in the sense that it does not depend on particular channel coding and modulation schemes used [36]. The outage probability achieved by FD-LE is given by $\mathrm{P}(\log(1+\mathrm{SINR}_{0,kl}^{\mathrm{LE}}) < R_0)$, where R_i, $0 \le i \le M$, denotes U_i's target data rate. It is difficult to analyze the outage probability for the following two reasons. First, the $D_0^{k,l}$, $k \in \{0,\cdots,N-1\}$, $l \in \{0,\cdots,M-1\}$, are not statistically independent, and second, the distribution of a sum of the inverse of exponentially distributed random variables is difficult to characterize. The following lemma provides an asymptotic result for the outage probability based on the SINR provided in Lemma 1.

Lemma 2: If $\gamma_0^2 > \gamma_1^2\epsilon_0$, the diversity order achieved by FD-LE is one, where $\epsilon_0 = 2^{R_0} - 1$. Otherwise, the outage probability is always one.

Proof: Please refer to Appendix C. □

Remark 3: Recall that the diversity order achieved by OTFS-OMA, where the high-mobility user, U_0, solely occupies the bandwidth resources, is also one. Therefore, the use of OTFS-NOMA ensures that the additional M low-mobility users are served without compromising U_0's diversity order, which improves the spectral efficiency compared to OTFS-OMA.

B. Design and Performance of FD-DFE

Different from FD-LE, which is a linear equalizer, FD-DFE is based on the idea of feeding back previously detected symbols. Since both $\mathbf{x}_0$ and $\mathbf{x}_q$, $q \ge 1$, experience the same fading channel, we first define $\mathbf{x} = \gamma_0\mathbf{x}_0 + \sum_{q=1}^{M}\gamma_q\mathbf{x}_q$, which are the signals to be recovered by FD-DFE. Given the received signal vector shown in (12), the outputs of FD-DFE are given by

$$\hat{\mathbf{x}} = \mathbf{P}_0\mathbf{y}_0 - \mathbf{G}_0\check{\mathbf{x}}, \quad (20)$$

where $\check{\mathbf{x}}$ contains the decisions made on the symbols $\mathbf{x}$, $\mathbf{P}_0$ is the feed-forward part of the equalizer, and $\mathbf{G}_0$ is the feedback part of the equalizer. Similar to [34], [35], we use the following choices for $\mathbf{P}_0$ and $\mathbf{G}_0$: $\mathbf{P}_0 = \mathbf{L}_0(\mathbf{H}_0^H\mathbf{H}_0)^{-1}\mathbf{H}_0^H$, $\mathbf{G}_0 = \mathbf{L}_0 - \mathbf{I}_{NM}$, where $\mathbf{L}_0$ is a lower triangular matrix with its main diagonal elements being ones in order to ensure causality of the feedback signals. With the above choices for $\mathbf{P}_0$ and $\mathbf{G}_0$, U_0's signals can be detected as follows:

$$\hat{\mathbf{x}} = \mathbf{L}_0(\mathbf{H}_0^H\mathbf{H}_0)^{-1}\mathbf{H}_0^H\mathbf{y}_0 - (\mathbf{L}_0 - \mathbf{I}_{NM})\check{\mathbf{x}}. \tag{21}$$

For FD-DFE, $\mathbf{L}_0$ is obtained from the Cholesky decomposition of $\mathbf{H}_0$, i.e., $\mathbf{H}_0^H\mathbf{H}_0 = \mathbf{L}_0^H\boldsymbol{\Lambda}_0\mathbf{L}_0$, where $\mathbf{L}_0$ is the desirable lower triangular matrix, and $\boldsymbol{\Lambda}_0$ is a diagonal matrix. Therefore, the estimates of $\mathbf{x}_0$ can be rewritten as follows:

$$\hat{\mathbf{x}} = \mathbf{x} + \mathbf{L}_0(\mathbf{H}_0^H\mathbf{H}_0)^{-1}\mathbf{H}_0^H\mathbf{z}_0 \tag{22}$$

$$= \gamma_0\mathbf{x}_0 + \underbrace{\sum_{q=1}^{M}\gamma_q\mathbf{x}_q + \mathbf{L}_0(\mathbf{H}_0^H\mathbf{H}_0)^{-1}\mathbf{H}_0^H\mathbf{z}_0}_{\text{Interference and noise terms}}, \tag{23}$$

where perfect decision-making is assumed, i.e., $\check{\mathbf{x}} = \mathbf{x}$, and there is no error propagation [35], [37], [38]. We note that (23) yields an upper bound on the reception reliability of FD-DFE when error propagation cannot be completely avoided.

Following steps similar to those in the proof of Lemma 1, the covariance matrix for the interference-plus-noise term can be found as follows:

$$\mathbf{C}_{\text{cov}} = \rho\gamma_1^2\mathbf{I}_{MN} + \mathbf{L}_0(\mathbf{H}_0^H\mathbf{H}_0)^{-1}\mathbf{L}_0^H = \rho\gamma_1^2\mathbf{I}_{MN} + \boldsymbol{\Lambda}_0^{-1}, \tag{24}$$

where the last step follows from the fact that $\mathbf{L}_0$ is obtained from the Cholesky decomposition of $\mathbf{H}_0$. Therefore, the SINR for detecting $x_0[k,l]$ can be expressed as follows:

$$\mathrm{SINR}_{0,kl} = \frac{\rho\gamma_0^2}{\rho\gamma_1^2 + \lambda_{0,kl}^{-1}}, \tag{25}$$

where $\lambda_{0,kl}$ is the $(kM+l+1)$-th element on the main diagonal of $\boldsymbol{\Lambda}_0$.

Remark 4: We note that there is a fundamental difference between the two equalization schemes. One can observe from (19) that the SINRs achieved by FD-LE for different $x_0[k,l]$ are identical. However, for FD-DFE, different symbols experience different effective fading gains, $\lambda_{0,kl}$. Therefore, FD-DFE can realize unequal error protection for data streams with different priorities. This comes at the price of a higher computational complexity.

We further note that the use of FD-DFE also ensures that multi-path diversity can be harvested, as shown in the following. The outage performance analysis for FD-DFE requires knowledge of the distribution of the effective channel gains, $\lambda_{0,kl}$. Because of the implicit relationship between $\boldsymbol{\Lambda}_0$ and $\mathbf{H}_0$, a general expression for the outage probability achieved by FD-DFE is difficult to obtain. However, analytical results can be developed for special cases to show that the use of FD-DFE can realize the maximal multi-path diversity.

In particular, the SINR for $x_0[N-1,M-1]$ is a function of $\lambda_{0,(N-1)(M-1)}$ which is the last element on the main diagonal of $\boldsymbol{\Lambda}_0$. Recall that $\boldsymbol{\Lambda}_0$ is obtained via Cholesky decomposition, i.e., $\mathbf{H}_0^H\mathbf{H}_0 = \mathbf{L}_0^H\boldsymbol{\Lambda}_0\mathbf{L}_0$. Because $\mathbf{L}_0$ is a lower triangular matrix, $\lambda_{0,(N-1)(M-1)}$ is equal to the element of $\mathbf{H}_0^H\mathbf{H}_0$ located in the NM-th column and the NM-th row, which means

$$\lambda_{0,(N-1)(M-1)} = \sum_{p=0}^{P_0}|h_{0,p}|^2. \tag{26}$$

Since the channel gains are i.i.d. and follow $h_{0,p} \sim CN(0,\frac{1}{P_0+1})$, the probability density function (pdf) of $\sqrt{P_0+1}\lambda_{0,(N-1)(M-1)}$ is given by

$$f(x) = \frac{1}{P_0!}e^{-x}x^{P_0}. \tag{27}$$

By using the above pdf, the outage probability and the diversity order can be obtained by some algebraic manipulations, as shown in the following corollary.

Corollary 1: Assume $\gamma_0^2 > \gamma_1^2\epsilon_0$. The use of FD-DFE realizes the following outage probability for detection of $x_0[N-1,M-1]$:

$$\mathrm{P}_{N-1,M-1}^0 = \frac{1}{P_0!}g\left(P_0+1, \frac{\epsilon_0(P_0+1)}{\rho(\gamma_0^2-\gamma_1^2\epsilon_0)}\right), \tag{28}$$

where $g(\cdot)$ denotes the incomplete Gamma function. The full multi-path diversity order, P_0+1, is achievable for $x_0[N-1, M-1]$.

Remark 5: The results in Corollary 1 can be extended to OTFS-OMA with FD-DFE straightforwardly. We also note that diversity gains larger than one are not achievable with FD-LE as shown in Lemma 2, which is one of the disadvantages of FD-LE compared to FD-DFE.

Remark 6: We note that not all NM data streams can benefit from the full diversity gain. The simulation results provided in Section VII (Fig. 2) show that the diversity orders achievable for $x_0[k,l]$, $k < N-1$ and $l < M-1$, are smaller than that for $x_0[N-1,M-1]$, and the diversity order for $x_0[0,0]$ is one, i.e., the same value as for FD-LE. We further note that the diversity result in Corollary 1 is obtained by assuming that there is no error propagation, i.e., it is assumed that when detecting the i-th element of $\mathbf{x}$ in (21), the first $(i-1)$ elements of $\mathbf{x}$ have already been correctly detected. Because of this assumption, the diversity gain developed in Corollary 1 is an upper bound on the diversity gain achieved by FD-DFE. If the assumption does not hold, the diversity orders for $x_0[k,l]$ will be capped by the worst case, i.e., the diversity gain for $x_0[0,0]$ which is one.

Remark 7: FD-DFE entails a higher implementation complexity than FD-LE, as explained in the following. The complexity of FD-LE is mainly caused by computing the inversion of $\mathbf{H}_0^H\mathbf{H}_0$. However, for FD-DFE, $\mathbf{L}_0$ needs to be computed, in addition to $(\mathbf{H}_0^H\mathbf{H}_0)^{-1}$, as shown in (21). Recall that $\mathbf{L}_0$ is obtained from the Cholesky decomposition of the $NM \times NM$ matrix $\mathbf{H}_0$, which entails a computational complexity of $\mathcal{O}(N^3M^3)$. Therefore, the computational complexity of FD-DFE is higher than that of FD-LE, but FD-DFE offers a performance gain in terms of reception reliability compared to FD-LE, as shown in Section VII.

V. Downlink OTFS-NOMA - Detecting the NOMA Users' Signals

Successive interference cancellation (SIC) will be carried out by the NOMA users, where each NOMA user first decodes the high mobility user's signal in the delay-Doppler plane and then decodes its own signal in the time-frequency plane. The two stages of SIC are discussed in the following two subsections, respectively.

A. Stage I of SIC

Following steps similar to the ones in the previous section, each NOMA user also observes the mixture of the $(M+1)$ users' signals in the delay-Doppler plane as follows:

$$\mathbf{y}_i = \gamma_0 \mathbf{H}_i \mathbf{x}_0 + \underbrace{\sum_{q=1}^{M} \gamma_q \mathbf{H}_i \mathbf{x}_q + \mathbf{z}_i}_{\text{Interference and noise terms}} , \tag{29}$$

where $\mathbf{H}_i$ and $\mathbf{z}_i$ are defined similar to $\mathbf{H}_0$ and $\mathbf{z}_0$, respectively.

We assume that the low-mobility NOMA users do not experience Doppler shift, and therefore, their channels can be simplified as follows:

$$h_i(\tau) = \sum_{p=0}^{P_i} h_{i,p} \delta(\tau - \tau_{i,p}), \tag{30}$$

for $1 \leq i \leq K$, which means that each NOMA user's channel matrix, $\mathbf{H}_i$, $1 \leq i \leq N$, is a block-diagonal matrix, i.e., $\mathbf{A}_{i,0}$ is a non-zero circulant matrix and $\mathbf{A}_{i,n} = \mathbf{0}_{M\times M}$, for $1 \leq n \leq N-1$. Therefore, each NOMA user can divide its observation vector into N equal-length sub-vectors, i.e., $\mathbf{y}_i = \left[\mathbf{y}_{i,0}^T \cdots \mathbf{y}_{i,N-1}^T\right]^T$, which yields the following simplified system model:

$$\mathbf{y}_{i,n} = \gamma_0 \mathbf{A}_{i,0} \mathbf{x}_{0,n} + \sum_{q=1}^{M} \gamma_q \mathbf{A}_{i,0} \mathbf{x}_{q,n} + \mathbf{z}_{i,n}, \tag{31}$$

where, similar to $\mathbf{y}_{i,n}$, $\mathbf{x}_{i,n}$ and $\mathbf{z}_{i,n}$ are obtained from $\mathbf{x}_i$ and $\mathbf{z}_i$, respectively. Therefore, unlike the high-mobility user, the NOMA users can perform their signal detection based on reduced-size observation vectors, which reduces the computational complexity.

Since $\mathbf{A}_{i,0}$ is a circulant matrix, the two equalization approaches used in the previous section are still applicable. First, we consider the use of FD-LE. Following the same steps as in the proof for Proposition 1, in the first step of FD-LE, the DFT matrix is applied to the reduced-size observation vector, which yields the following:

$$\tilde{\mathbf{y}}_{i,n} = \tilde{\mathbf{D}}_i \mathbf{F}_M^H \left(\gamma_0 \mathbf{x}_{0,n} + \sum_{q=1}^{M} \gamma_q \mathbf{x}_{q,n} \right) + \tilde{\mathbf{z}}_{i,n}, \tag{32}$$

where $\tilde{\mathbf{y}}_{i,n} = \mathbf{F}_M^H \mathbf{y}_{i,n}$ and $\tilde{\mathbf{z}}_{i,n} = \mathbf{F}_M^H \mathbf{z}_{i,n}$. Compared to $\mathbf{D}_i$ in Proposition 1 which is an $NM \times NM$ matrix, $\tilde{\mathbf{D}}_i$ is an $M \times M$ diagonal matrix, and its $(l+1)$-th main diagonal element is given by $\tilde{D}_i^l = \sum_{m=0}^{M-1} a_{i,0}^{m,1} e^{j2\pi\frac{lm}{M}}$, for $0 \leq l \leq M-1$, where $a_{i,0}^{m,1}$ is the element located in the $(m+1)$-th row and the first column of $\mathbf{A}_{i,0}$. Unlike conventional OFDM, which uses $\mathbf{F}_M$ at the receiver, $\mathbf{F}_M^H$ is used here. Because $\mathbf{F}_M^H \mathbf{A}_{i,0} \mathbf{F}_M = \left[\mathbf{F}_M \mathbf{A}_{i,0}^* \mathbf{F}_M^H\right]^*$, the sign of the exponent of the exponential component of $\tilde{D}_i^l$ is different from that in the conventional case.

In the second step of FD-LE, $\mathbf{F}_M \tilde{\mathbf{D}}_i^{-1}$ is applied to $\tilde{\mathbf{y}}_{i,n}$. Following steps similar to the ones in the proof for Lemma 1, the SINR for detecting $x_0[k,l]$ can be obtained as follows:

$$\mathrm{SINR}_{0,kl}^{i,\mathrm{LE}} = \frac{\rho\gamma_0^2}{\rho\gamma_1^2 + \frac{1}{M}\sum_{\tilde{l}=0}^{M-1} |\tilde{D}_i^{\tilde{l}}|^{-2}}. \tag{33}$$

We note that $\mathrm{SINR}_{0,k_1 l}^{i,\mathrm{LE}} = \mathrm{SINR}_{0,k_2 l}^{i,\mathrm{LE}}$, for $k_1 \neq k_2$, due to the time invariant nature of the channels.

If FD-DFE is used, the corresponding SINR for detecting $x_0[k,l]$ is given by

$$\mathrm{SINR}_{0,kl}^{i,\mathrm{DFE}} = \frac{\rho\gamma_0^2}{\rho\gamma_1^2 + \tilde{\lambda}_{0,l}^{-1}}, \tag{34}$$

where $\tilde{\lambda}_{0,l}$ is obtained from the Cholesky decomposition of $\mathbf{A}_{i,0}$. The details for the derivation of (34) are omitted here due to space limitations.

B. Stage II of SIC

Assume that U_0's NM signals can be decoded and removed successfully, which means that, in the time-frequency plane, the NOMA users observe the following:

$$\begin{aligned} Y_i[n,m] &= \sum_{q=1}^{M} \gamma_q H_i[n,m] X_q[n,m] + W_i[n,m] \\ &= \gamma_1 H_i[n,m] x_{m+1}(n) + W_i[n,m], \end{aligned} \tag{35}$$

where the last step follows from the mapping scheme used in (6) and it is assumed that all NOMA users employ the same power allocation coefficient. We note that U_i is only interested in $Y_i[n,i-1]$, $0 \leq n \leq N-1$. Therefore, U_i's n-th information bearing signal, $x_i(n)$, can be detected by applying a one-tap equalizer as follows:

$$\hat{x}_i(n) = \frac{Y_i[n,i-1]}{\gamma_1 H_i[n,i-1]}, \tag{36}$$

which means that the SNR for detecting $x_i(n)$ is given by

$$\mathrm{SNR}_{i,n} = \rho\gamma_1^2 |\tilde{D}_i^{i-1}|^2, \tag{37}$$

since $W_i[n,i-1]$ is white Gaussian noise and $H_i[n,i-1] = \tilde{D}_i^{i-1}$. We note that $\mathrm{SNR}_{i,n_1} = \mathrm{SNR}_{i,n_2}$, for $n_1 \neq n_2$, which is due to the time-invariant nature of the channel.

Without loss of generality, assume that the same target data rate R_i is used for $x_i(n)$, $0 \leq n \leq N-1$. Therefore, the outage probability for $x_i(n)$ is given by

$$\begin{aligned} &\mathrm{P}_{i,n}^{\mathrm{LE}} \\ &= 1 - \mathrm{P}\left(\mathrm{SNR}_{i,n} > \epsilon_i, \mathrm{SINR}_{0,kl}^{i,\mathrm{LE}} > \epsilon_0, \forall l\right) \\ &= 1 - \mathrm{P}\left(\rho\gamma_1^2 |\tilde{D}_i^{i-1}|^2 > \epsilon_i, \frac{\rho\gamma_0^2}{\rho\gamma_1^2 + \frac{1}{M}\sum_{l=0}^{M-1} |\tilde{D}_i^l|^{-2}} > \epsilon_0\right), \end{aligned} \tag{38}$$

if FD-LE is used in the first stage of SIC. If FD-DFE is used in the first stage of SIC, the outage probability for $x_i(n)$ is given by

$$\begin{aligned} \mathrm{P}_{i,n}^{\mathrm{DFE}} &= 1 - \mathrm{P}\left(\mathrm{SNR}_{i,n} > \epsilon_i, \mathrm{SINR}_{0,kl}^{i,\mathrm{DFE}} > \epsilon_0, \forall l\right) \\ &= 1 - \mathrm{P}\left(\rho\gamma_1^2 |\tilde{D}_i^{i-1}|^2 > \epsilon_i, \frac{\rho\gamma_0^2}{\rho\gamma_1^2 + \tilde{\lambda}_{0,l}^{-1}} > \epsilon_0, \forall l\right), \end{aligned} \tag{39}$$

where $\epsilon_i = 2^{R_i} - 1$. Again because of the correlation between the random variables $|\tilde{D}_i^l|^{-2}$ and $\tilde{\lambda}_{0,l}$, the exact expressions for the outage probabilities are difficult to obtain. Alternatively, the achievable diversity order is analyzed in the following subsections.

1) Random User Scheduling: If the M users are randomly selected from the K available users, which means that each $|\tilde{D}_i^l|^2$ is complex Gaussian distributed. For the FD-LE case, the outage probability, $\mathrm{P}_{i,n}^{\mathrm{LE}}$, can be upper bounded as follows:

$$\mathrm{P}_{i,n}^{\mathrm{LE}} \leq 1 - \mathrm{P}\left(\rho\gamma_1^2|\tilde{D}_i^{\min}|^2 > \epsilon_i, \frac{\rho\gamma_0^2}{\rho\gamma_1^2 + |\tilde{D}_i^{\min}|^{-2}} > \epsilon_0\right), \tag{40}$$

where $|\tilde{D}_i^{\min}|^2 = \min\{|\tilde{D}_i^m|^2, 0 \leq m \leq M-1\}$. The upper bound on the outage probability in (40) can be rewritten as follows:

$$\mathrm{P}_{i,n}^{\mathrm{LE}} \leq 1 - \mathrm{P}\left(|\tilde{D}_i^{\min}|^2 > \bar{\epsilon}\right), \tag{41}$$

where $\bar{\epsilon} = \max\left\{\frac{\epsilon_0}{\rho(\gamma_0^2 - \gamma_1^2\epsilon_0)}, \frac{\epsilon_i}{\rho\gamma_1^2}\right\}$. As a result, an upper bound on the outage probability can be obtained as follows:

$$\mathrm{P}_{i,n}^{\mathrm{LE}} \leq \mathrm{P}\left(|\tilde{D}_i^{\min}|^2 < \bar{\epsilon}\right) \leq M\mathrm{P}\left(|\tilde{D}_i^{0}|^2 < \bar{\epsilon}\right) \doteq \frac{1}{\rho}, \tag{42}$$

where $\mathrm{P}^o \doteq \rho^{-d}$ denotes exponential equality, i.e., $d = -\lim\limits_{\rho\to\infty} \frac{\log \mathrm{P}^o}{\log \rho}$ [36]. Therefore, the following corollary can be obtained.

Corollary 2: For random user scheduling and FD-LE, a diversity order of 1 is achievable at the NOMA users.

Our simulation results in Section VII show that a diversity order of 1 is also achievable for FD-DFE, although we do not have a formal proof for this conclusion, yet.

2) Realizing Multi-User Diversity: The diversity order of OTFS-NOMA can be improved by carrying out opportunistic user scheduling, which yields multi-user diversity gains. For illustration purpose, we propose a greedy user scheduling policy, where a single NOMA user is scheduled to transmit in all resource blocks of the time-frequency plane. From the analysis of the random scheduling case we deduce that $|\tilde{D}_i^{\min}|^2$ is critical to the outage performance. Therefore, the scheduled NOMA user, denoted by U_{i^*}, is selected based on the following criterion:

$$i^* = \arg \max_{i\in\{1,\cdots,K\}} \left\{|\tilde{D}_i^{\min}|^2\right\}. \tag{43}$$

By using the assumption that the users' channel gains are independent and following steps similar to the ones in the proof for Lemma 2, the following corollary can be obtained in a straightforward manner.

Corollary 3: For FD-LE, the user scheduling strategy shown in (43) *realizes the maximal multi-user diversity gain, K.*

Remark 8: The reason why a multi-user diversity gain of K can be realized by the proposed scheduling strategy is explained in the following. Recall that the SINR for FD-LE to detect $x_0[k,l]$ is $\mathrm{SINR}_{0,kl}^{i,\mathrm{LE}} = \frac{\rho\gamma_0^2}{\rho\gamma_1^2 + \frac{1}{M}\sum_{\tilde{l}=0}^{M-1}|\tilde{D}_i^{\tilde{l}}|^{-2}}$. If this SINR is too small, the first stage of SIC will fail and an outage event will occur. To improve the SINR, it is important to ensure that for a scheduled user, its weakest channel gain, $|\tilde{D}_i^{\min}|^2 = \min\{|\tilde{D}_i^m|^2, 0 \leq m \leq M-1\}$, is not too small. The used scheduling strategy shown in (43) is essentially a max-min strategy and ensures that the user with the strongest $|\tilde{D}_i^{\min}|^2$ is selected from the K candidates, which effectively exploits multi-user diversity.

We note that the user scheduling strategy shown in (43) is also useful for improving the performance of FD-DFE, as shown in Section VII.

VI. Uplink OTFS-NOMA Transmission

The design of uplink OTFS-NOMA is similar to that of downlink OTFS-NOMA, and due to space limitations, we mainly focus on the difference between the two cases in this section. Again, we assume that U_0 is grouped with M NOMA users, selected from the K available users. U_0's NM signals are placed in the delay-Doppler plane, and are denoted by $x_0[k,l]$, where $0 \leq k \leq N-1$ and $0 \leq l \leq M-1$. The corresponding time-frequency signals, $X_0[n,m]$, are obtained by applying ISFFT to $x_0[k,l]$. On the other hand, the NOMA users' signals, $x_i(n)$, are mapped to time-frequency signals, $X_i[n,m]$, according to (6).

Following steps similar to the ones for the downlink case, the base station's observations in the time-frequency plane are given by

$$\begin{aligned} Y[n,m] &= \sum_{q=0}^{M} H_q[n,m]X_q[n,m] + W[n,m] \\ &= \frac{H_0(n,m)}{NM}\sum_{k=0}^{N-1}\sum_{l=0}^{M-1} x_0[k,l]e^{j2\pi\left(\frac{kn}{N}-\frac{ml}{M}\right)} \\ &\quad + \sum_{q=1}^{M} H_q[n,m]X_q[n,m] + W[n,m], \end{aligned} \tag{44}$$

where $W[n,m]$ is the Gaussian noise at the base station in the time-frequency plane. We assume that all users employ the same transmit pulse as well as the same transmit power. The base station applies SIC to first detect the NOMA users' signals in the time-frequency plane, and then tries to detect the high-mobility user's signals in the delay-Doppler plane, as shown in the following two subsections.

A. Stage I of SIC

The base station will first try to detect the NOMA users' signals in the time-frequency plane by treating the signals from U_0 as noise, which is the first stage of SIC.

By using (6), $x_i(n)$ can be estimated as follows:

$$\begin{aligned} \hat{x}_i(n) &= \frac{Y[n,i-1]}{H_i[n,i-1]} \\ &= x_i[n] + \frac{H_0[n,i-1]X_0[n,i-1] + W[n,i-1]}{H_i[n,i-1]}. \end{aligned} \tag{45}$$

Define an $NM \times 1$ vector, $\bar{\mathbf{x}}_0$, whose $(nM+m+1)$-th element is $X_0[n,m]$. Recall that $X_0[n,m]$ is obtained from the ISFFT of $x_0[k,l]$, i.e.,

$$\bar{\mathbf{x}}_0 = (\mathbf{F}_N^H \otimes \mathbf{F}_M)\mathbf{x}_0, \tag{46}$$

which means $X_0[n, m]$ follows the same distribution as $x_0[k, l]$. By applying steps similar to those in the proof for Lemma 1, the SINR for detecting $x_i(n)$ is given by

$$\mathrm{SINR}_{i,n} = \frac{\rho|H_i[n, i-1]|^2}{\rho|H_0[n, i-1]|^2 + 1}. \tag{47}$$

Unlike downlink OTFS-NOMA, there are two possible strategies for uplink OTFS-NOMA to combat multiple access interference, as shown in the following two subsections.

1) Adaptive-Rate Transmission: One strategy to combat multiple access interference is to impose the following constraint on $x_i(n)$:

$$R_{i,n} \leq \log\left(1 + \frac{\rho|H_i[n, i-1]|^2}{\rho|H_0[n, i-1]|^2 + 1}\right), \tag{48}$$

which means that the first stage of SIC is guaranteed to be successful. Therefore, the M low-mobility users are served without affecting U_0's outage probability, i.e., the use of NOMA is transparent to U_0.

Because U_i's data rate is adaptive, outage events when decoding $x_i(n)$ do not happen, which means that an appropriate criterion for the performance evaluation is the ergodic rate. Recall that $H_i[n, i-1] = \tilde{D}_i^{i-1}$ and $H_0[n, i-1] = D_0^{n,i-1}$. Therefore, U_i's ergodic rate is given by

$$\mathcal{E}\{R_{i,n}\} \leq \mathcal{E}\left\{\log\left(1 + \frac{\rho|\tilde{D}_i^{i-1}|^2}{\rho|D_0^{n,i-1}|^2 + 1}\right)\right\}. \tag{49}$$

We note that the ergodic rate of uplink OTFS-NOMA can be further improved by modifying the user scheduling strategy proposed in (43), as shown in the following. Particularly, denote the NOMA user which is scheduled to transmit in the m-th frequency subchannel by $\mathrm{U}_{i_m^*}$, and this user is selected by using the following criterion:

$$i_m^* = \arg \max_{i \in \{1, \cdots, K\}} \left\{|\tilde{D}_i^m|^2\right\}. \tag{50}$$

We note that a single user might be scheduled on multiple frequency channels, which reduces user fairness.

Because the integration of the logarithm function appearing in (49) leads to non-insightful special functions, we will use simulations to evaluate the ergodic rate of OTFS-NOMA in Section VII.

2) Fixed-Rate Transmission: If the NOMA users do not have the capabilities to adapt their transmission rates, they have to use fixed data rates R_i for transmission, which means that outage events can happen and the achieved outage performance is analyzed in the following. For illustration purposes, we focus on the case when the user scheduling strategy shown in (50) is used.

The outage probability for detecting $x_{i_m^*}(n)$ is given by

$$\mathrm{P}_{i_m^*,n} = \mathrm{P}\left(\log\left(1 + \frac{\rho|\tilde{D}_{i_m^*}^{i_m^*-1}|^2}{\rho|D_0^{n,i_m^*-1}|^2 + 1}\right) < R_{i_m^*}\right). \tag{51}$$

Following steps similar to the ones in the proof for Lemma 2, we can show that $|\tilde{D}_{i_m^*}^{i_m^*-1}|^2$ and $|D_0^{n,i_m^*-1}|^2$ are independent, and the use of the user scheduling scheme in (50) simplifies the outage probability as follows:

$$\begin{aligned}\mathrm{P}_{i_m^*,n} &= \mathrm{P}\left(\log\left(1 + \frac{\rho|\tilde{D}_{i_m^*}^{i_m^*-1}|^2}{\rho|D_0^{n,i_m^*-1}|^2 + 1}\right) < R_{i_m^*}\right) \\ &= \int_0^\infty \left(1 - e^{-\frac{\epsilon_{i_m^*}(1+\rho y)}{\rho}}\right)^K e^{-y} dy, \end{aligned} \tag{52}$$

where we use the fact that the cumulative distribution function of $|\tilde{D}_{i_m^*}^{i_m^*-1}|^2$ is $(1 - e^{-x})^K$ because of the adopted user scheduling strategy.

The outage probability can be further simplified as follows:

$$\begin{aligned}\mathrm{P}_{i_m^*,n} &= \sum_{k=0}^{K} \binom{K}{k} (-1)^k \int_0^\infty e^{-\frac{k\epsilon_{i_m^*}(1+\rho y)}{\rho} - y} dy \\ &= \sum_{k=0}^{K} \binom{K}{k} (-1)^k e^{-\frac{k\epsilon_{i_m^*}}{\rho}} \frac{1}{k\epsilon_{i_m^*} + 1}. \end{aligned} \tag{53}$$

At high SNR, the outage probability can be approximated as follows:

$$\mathrm{P}_{i_m^*,n} \approx \sum_{k=0}^{K} \binom{K}{k} (-1)^k \frac{1}{k\epsilon_{i_m^*} + 1}, \tag{54}$$

which is no longer a function of ρ, i.e., the outage probability has an error floor at high SNR. This is due to the fact that $\mathrm{U}_{i_m^*}$ is subject to strong interference from U_0.

However, we can show that the error floor experienced by $\mathrm{U}_{i_m^*}$ can be reduced by increasing K, i.e., inviting more opportunistic users for NOMA transmission. In particular, assuming $K\epsilon_{i_m^*} \to 0$, the outage probability can be approximated as follows:

$$\begin{aligned}\mathrm{P}_{i_m^*,n} &\approx \sum_{k=0}^{K} \binom{K}{k} (-1)^k \left(1 + k\epsilon_{i_m^*}\right)^{-1} \\ &\approx \sum_{k=0}^{K} \binom{K}{k} (-1)^k \sum_{l=0}^{\infty} (-1)^l k^l \epsilon_{i_m^*}^l, \end{aligned} \tag{55}$$

where we use the fact that $(1+x)^{-1} = \sum_{l=0}^{\infty} (-1)^l x^l$, $|x| < 1$. Therefore, the error floor at high SNR can be approximated as follows:

$$\begin{aligned}\mathrm{P}_{i_m^*,n} &\approx \sum_{l=0}^{\infty} (-1)^l \epsilon_{i_m^*}^l \sum_{k=0}^{K} \binom{K}{k} (-1)^k k^l \\ &\approx (-1)^K \epsilon_{i_m^*}^K (-1)^K K! = K! \epsilon_{i_m^*}^K, \end{aligned} \tag{56}$$

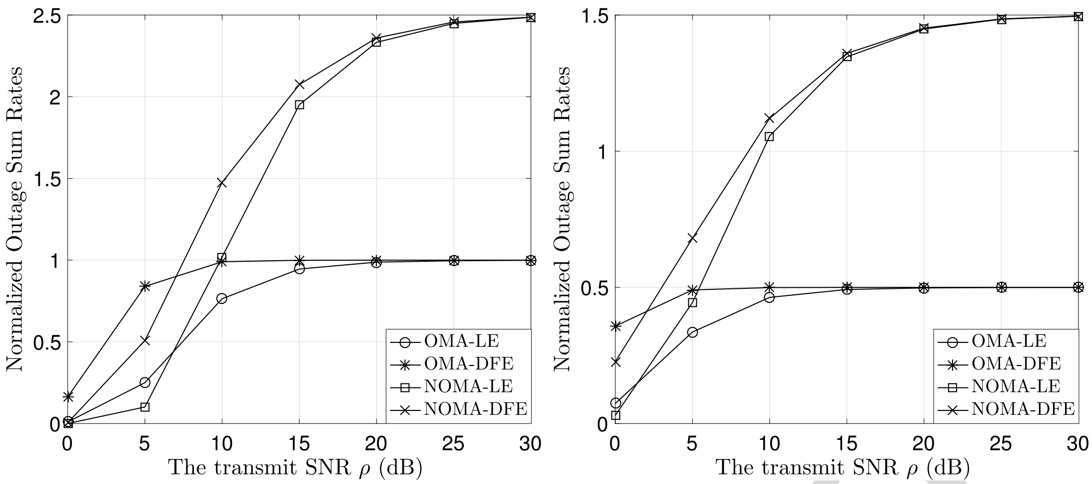
where we use the identities $\sum_{k=0}^{K} \binom{K}{k} (-1)^k k^l = 0$, for $l < K$ and $\sum_{k=0}^{K} \binom{K}{k} (-1)^k k^K = (-1)^K K!$.

The conclusion that increasing K reduces the error floor can be confirmed by defining $f(k) = k! \epsilon_{i_m^*}^k$ and using the following fact:

$$f(k) - f(k+1) = k! \epsilon_{i_m^*}^k \left(1 - (k+1)\epsilon_{i_m^*}\right) > 0, \tag{57}$$

where it is assumed that $k\epsilon_{i_m^*} \to 0$.

AQ:4



(a) $R_0 = 1$ BPCU and $R_i = 1.5$ BPCU (b) $R_0 = 0.5$ BPCU and $R_i = 1$ BPCU

Fig. 1. Impact of OTFS-NOMA on the downlink sum rates. $M = N = K = 16$. $P_0 = P_i = 3$. BPCU denotes bit per channel use. $\gamma_0^2 = \frac{3}{4}$ and $\gamma_i^2 = \frac{1}{4}$ for $i > 0$. Random user scheduling is used.

B. Stage II of SIC

If adaptive transmission is used, the NOMA users' signals can be detected successfully during the first stage of SIC. Therefore, they can be removed from the observations at the base station, i.e., $\bar{Y}[n,m] = Y[n,m] - \sum_{q=1}^{N} H_q(n,m)X_q[n,m]$, and SFFT is applied to obtain the delay-Doppler observations as follows:

$$\begin{aligned} y_0[k,l] &= \frac{1}{NM}\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}\bar{Y}[n,m]e^{-j2\pi\left(\frac{nk}{N}-\frac{ml}{M}\right)} \\ &= \sum_{p=0}^{P_0} h_{0,p}x_0[(k-k_{\mu_{0,p}})_N,(l-l_{\tau_{0,p}})_M]+z[k,l], \quad (58) \end{aligned}$$

where $z[k,l]$ denote additive noise. U_0's signals can be detected by applying either of the two considered equalization approaches, and the same performance as for OTFS-OMA can be realized. The analytical development is similar to the downlink case, and hence is omitted due to space limitations.

However, if fixed-rate transmission is used, the uplink outage events for decoding $x_0[k,l]$ are different from the downlink ones, as shown in the following. Particularly, the use of FD-LE yields the following SINR expression for decoding $x_0[k,l]$:

$$\mathrm{SINR}_{0,kl}^{\mathrm{LE}} = \frac{\rho}{\frac{1}{NM}\sum_{k=0}^{N-1}\sum_{l=0}^{M-1}|D_0^{k,l}|^{-2}}. \quad (59)$$

If FD-DFE is used, the SNR for detection of $x_0[k,l]$ is given by

$$\mathrm{SINR}_{0,kl}^{\mathrm{DFE}} = \rho\lambda_{0,kl}. \quad (60)$$

Therefore, the outage probability for detecting $x_0[k,l]$ is given by

$$\begin{aligned} \mathrm{P}_{kl} &= 1 - \mathrm{P}\left(\mathrm{SINR}_{0,kl}^{\mathrm{DFE/LE}} > \epsilon_0, \mathrm{SNR}_{i,n} > \epsilon_i \forall i,n\right) \\ &\geq 1 - \mathrm{P}\left(\mathrm{SNR}_{i,n} > \epsilon_i \forall i,n\right) \geq \mathrm{P}\left(\mathrm{SNR}_{1,0} < \epsilon_i\right). \end{aligned}$$

Since $\mathrm{P}\left(\mathrm{SNR}_{1,0} < \epsilon_i\right)$ has an error floor as shown in the previous subsection, the uplink outage probability for detection of U_0's signals does not go to zero even if $\rho \to \infty$, which is different from the downlink case. Therefore, if fixed-rate transmission is used, adding the M low-mobility users into the bandwidth, which would be solely occupied by U_0 in OTFS-OMA, improves connectivity but degrades U_0's performance.

TABLE I

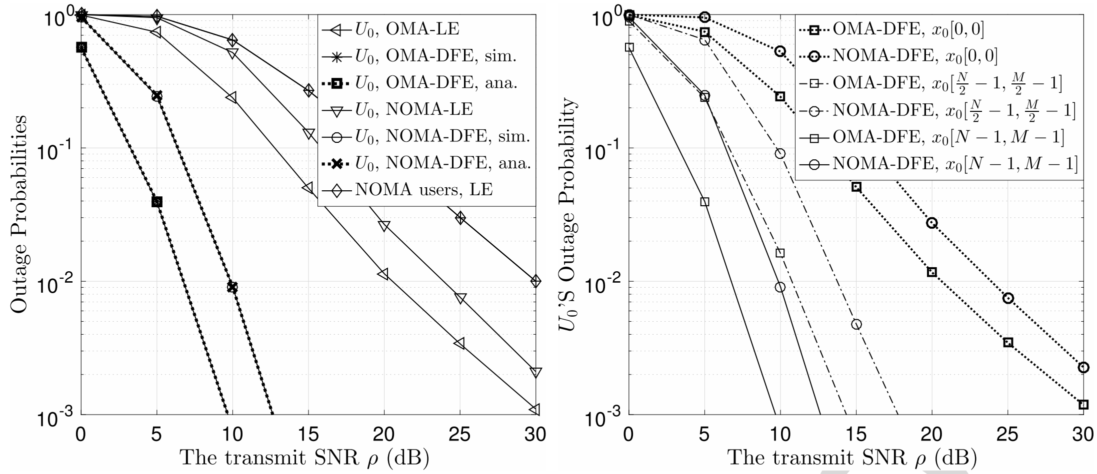
DELAY-DOPPLER PROFILE FOR U_0'S CHANNEL

Propagation path index (p)	0	1	2	3
Delay ($\tau_{0,p}$) μs	8.33	25	41.67	58.33
Delay tap index ($l_{\tau_{0,p}}$)	2	6	10	14
Doppler ($\nu_{0,p}$) Hz	0	0	468.8	468.8
Doppler tap index ($k_{\nu_{0,p}}$)	0	0	1	1

VII. NUMERICAL STUDIES

In this section, the performance of OTFS-NOMA is evaluated via computer simulations. Similar to [26]–[28], we first define the delay-Doppler profile for U_0's channel as shown in Table I, where $P_0 = 3$ and the subchannel spacing is $\Delta f = 7.5$ kHz. Therefore, the maximal speed corresponding to the largest Doppler shift $\nu_{0,3} = 468.8$ Hz is 126.6 km/h if the carrier frequency is $f_c = 4$ GHz. On the other hand, the NOMA users' channels are assumed to be time invariant with $P_i = 3$ propagation paths, i.e., $\tau_{i,p} = 0$ for $p \geq 4$, $i \geq 1$. For all the users' channels, we assume that $\sum_{p=0}^{P_i}\mathcal{E}\{|h_{i,p}|^2\} = 1$ and $|h_{i,p}|^2 \sim CN\left(0, \frac{1}{P_i+1}\right)$. For the fixed rate transmission scheme, a simple choice for power allocation ($\gamma_0^2 = \frac{3}{4}$ and $\gamma_i^2 = \frac{1}{4}$ for $i > 0$) is considered. The performance of OTFS-NOMA could be further improved by optimizing γ_i according to the users' channel conditions and QoS requirements.

In Fig. 1, downlink OTFS-NOMA transmission is evaluated by using the normalized outage sum rate as the performance criterion which is defined as $\frac{1}{NM}\sum_{k=0}^{N-1}\sum_{l=0}^{M-1}(1-\mathrm{P}_{0,kl})R_0$



(a) Outage probabilities of U_0 and the NOMA users (b) Performance of FD-DFE

Fig. 2. The outage performance of downlink OTFS-OMA and OTFS-NOMA. $M = N = K = 16$. $P_0 = P_i = 3$. $\gamma_0^2 = \frac{3}{4}$ and $\gamma_i^2 = \frac{1}{4}$ for $i > 0$. $R_0 = 0.5$ BPCU and $R_i = 1$ BPCU. In Fig. 2(a), for FD-DFE, the performance of $x_0[N-1, M-1]$ is shown. Random user scheduling is used.

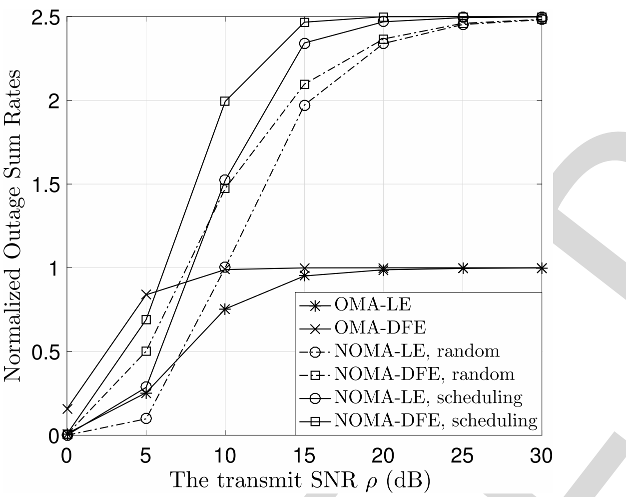


Fig. 3. Impact of user scheduling on the downlink outage sum rates. $P_0 = P_i = 3$. $R_0 = 1$ BPCU and $R_i = 1.5$ BPCU. $M = N = K = 16$, $\gamma_0^2 = \frac{3}{4}$ and $\gamma_i^2 = \frac{1}{4}$ for $i > 0$.

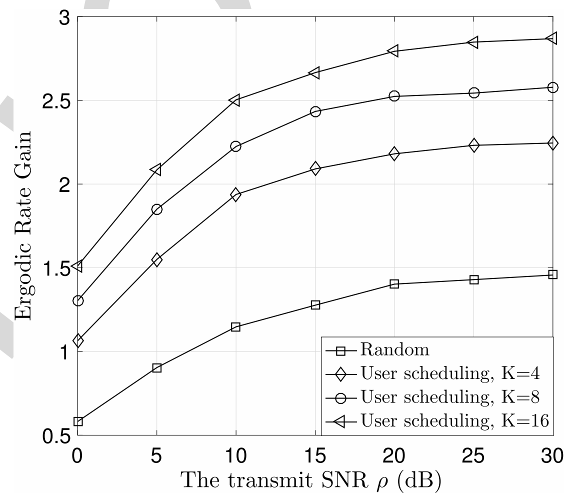
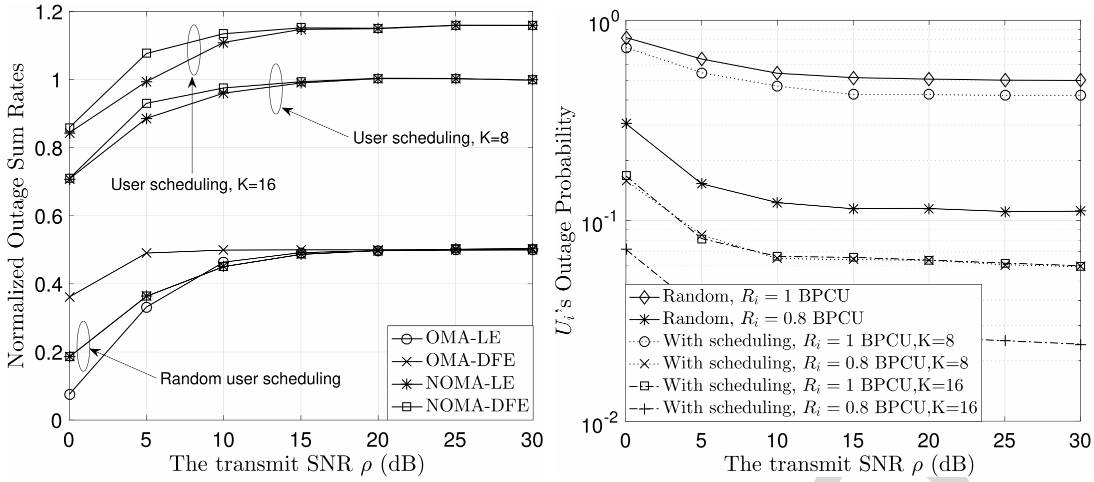


Fig. 4. The ergodic rate gain of OTFS-NOMA over OTFS-OMA. The NOMA users adapt their data rates according to (48). $P_0 = P_i = 3$. $M = N = 16$.

and $\frac{1}{NM}\sum_{k=0}^{N-1}\sum_{l=0}^{M-1}(1-\mathrm{P}_{0,kl})R_0 + \frac{1}{NM}\sum_{i=1}^{M}\sum_{n=0}^{N-1}(1-\mathrm{P}_{i,n})R_i$ for OTFS-OMA and OTFS-NOMA, respectively. Fig. 1 shows that the use of OTFS-NOMA can significantly improve the sum rate at high SNR for both considered choices of R_0 and R_i. The reason for this performance gain is the fact that the maximal sum rate achieved by OTFS-OMA is capped by R_0, whereas OTFS-NOMA can provide sum rates up to $R_0 + R_i$. Comparing Fig. 1(a) to Fig. 1(b), one can observe that the performance loss of OTFS-NOMA at low SNR can be mitigated by reducing the target data rates, since reducing the target rates improves the probability of successful SIC. Furthermore, both figures show that FD-DFE outperforms FD-LE in the entire considered range of SNRs; however, we note that the performance gain of FD-DFE over FD-LE is achieved at the expense of increased computational complexity.

In Fig. 2, the outage probabilities achieved by downlink OTFS-OMA and OTFS-NOMA are shown. As can be seen from Fig. 2(a), the diversity order achieved with FD-LE for detection of $x_0[k,l]$ is one, as expected from Lemma 2. As discussed in Section IV-B, one advantage of FD-DFE over FD-LE is that FD-DFE facilitates multi-path fading diversity gains, whereas FD-LE is limited to a diversity gain of one. This conclusion is confirmed by Fig. 2(a), where the analytical results developed in Corollary 1 are also verified. Fig. 2(b) shows the outage probabilities achieved by FD-DFE for different $x_0[k,l]$. As shown in the figure, the lowest outage probability is obtained for $x_0[N-1, M-1]$, whereas the outage probability of $x_0[0,0]$ is the largest, which is due to the fact that, in FD-DFE, different signals $x_0[k,l]$ are affected by different effective channel gains, $\lambda_{0,kl}$. Another important observation from the figures is that the FD-LE outage probability is the same as the FD-DFE outage probability for detection of $x_0[0,0]$, which fits the intuition that for FD-DFE the reliability of the first decision ($x_0[0,0]$) is the same as that of FD-LE. For the same reason, FD-LE and FD-DFE yield similar performance for detection of the NOMA users'



(a) Outage Sum Rate ($R_i = 1$ BPCU)

(b) Outage Probability

Fig. 5. The performance of uplink OTFS-NOMA. Fixed-rate transmission is used by the NOMA users. $M = N = 16$. $P_0 = P_i = 3$. $R_0 = 0.5$ BPCU. $\gamma_0^2 = \frac{3}{4}$ and $\gamma_i^2 = \frac{1}{4}$ for $i > 0$.

signals, since the FD-DFE outage performance is dominated by the reliability for detection of $x_0[0, 0]$, and hence is the same as that of FD-LE.

In addition to multi-path diversity, another degree of freedom available in the considered OTFS-NOMA downlink scenario is multi-user diversity, which can be harvested by applying user scheduling as discussed in Section V-B. Fig. 3 demonstrates the benefits of exploiting multi-user diversity. With random user scheduling, at low SNR, the performance of OTFS-NOMA is worse than that of OTFS-OMA, which is also consistent with Fig. 1. By increasing the number of users participating in OTFS-NOMA, the performance of OTFS-NOMA can be improved, particularly at low and moderate SNR. For example, for FD-LE, the performance of OTFS-NOMA approaches that of OTFS-OMA at low SNR by exploiting multi-user diversity, and for FD-DFE, an extra gain of 0.5 BPCU can be achieved at moderate SNR.

In Figs. 4 and 5, the performance of uplink OTFS-NOMA is evaluated. As discussed in Section VI, the NOMA users have two choices for their transmission rates, namely adaptive and fixed rate transmission. The use of adaptive rate transmission can ensure that the implementation of NOMA is transparent to U_0, which means that U_0's QoS requirements are strictly guaranteed. Since U_0 achieves the same performance for OTFS-NOMA and OTFS-OMA when adaptive rate transmission is used, we only focus on the NOMA users' performance, where the ergodic rate in (49) is used as the criterion. We note that this ergodic rate is the net performance gain of OTFS-NOMA over OTFS-OMA, which is the reason why the vertical axis in Fig. 4 is labeled 'Ergodic Rate Gain'. When the M users are randomly selected from the K NOMA users, the ergodic rate gain is moderate, e.g., 1.5 bit per channel use (BPCU) at $\rho = 30$ dB. By applying the scheduling strategy proposed in (50), the ergodic rate gain can be significantly improved, e.g., nearly by a factor of two compared to the random case with $K = 16$ and $\rho = 30$ dB.

Fig. 5 focuses on the case with fixed rate transmission, and similar to Fig. 1, the normalized outage sum rate is used as performance criterion in Fig. 5(a). One can observe that with random user scheduling, the sum rate of OTFS-NOMA is similar to that of OTFS-OMA. This is due to the fact that no interference mitigation strategy, such as power or rate allocation, is used for NOMA uplink transmission, which means that U_0 and the NOMA users cause strong interference to each other and SIC failure may happen frequently. By applying the user scheduling strategy proposed in (50), the channel conditions of the scheduled users become quite different, which facilitates the implementation of SIC. This benefit of user scheduling can be clearly observed in Fig. 5(a), where NOMA achieves a significant gain over OMA although advanced power or rate allocation strategies are not used. Fig. 5(a) also shows that the difference between the performance of FD-LE and FD-DFE is insignificant for the uplink case. This is due to the fact that the outage events during the first stage of SIC dominate the outage performance, and they are not affected by whether FD-LE or FD-DFE is employed. Another important observation from Fig. 5(a) is that the maximal sum rate $R_0 + R_i$ cannot be realized, even at high SNR. The reason for this behaviour is the existence of the error floor for the NOMA users' outage probabilities, as shown in Fig. 5(b). The analytical results provided in Section V-B show that increasing K can reduce the error floor, which is confirmed by Fig. 5(b).

VIII. Conclusion

In this paper, we have proposed OTFS-NOMA uplink and downlink transmission schemes, where users with different mobility profiles are grouped together for the implementation of NOMA. The analytical results developed in the paper demonstrate that both the high-mobility and the low-mobility users benefit from the application of OTFS-NOMA. In particular, the use of NOMA enables the spreading of the signals of a high-mobility user over a large amount of time-frequency resources, which enhances the OTFS resolution and improves the detection reliability. In addition, OTFS-NOMA ensures that the low-mobility users have access to the bandwidth resources which would be solely occupied by the high-mobility users in OTFS-OMA. Hence, OTFS-NOMA improves the spectral efficiency and reduces latency. An interesting topic for future works is studying the impact of non-zero fractional delays and fractional Doppler shifts on the performance of the developed OTFS-NOMA protocol. Furthermore, in this paper, the users' channel gains (the taps of the delay-Doppler impulse response) have been assumed to be Gaussian distributed, and an important direction for future research is to investigate the impact of other types of channel distributions on the performance of OTFS-NOMA. Moreover, the combination of emerging spectrally efficient 5G solutions, such as 5G New Radio Bandwidth Part (5G-NR-BWP) [39], [40] and software-controlled metasurfaces [41], with OTFS-NOMA is also a promising topic for future research.

Appendix A
Proof for Proposition 1

Intuitively, the use of $\mathbf{F}_N \otimes \mathbf{F}_M^H$ is analogous to the application of the ISFFT which transforms signals from the delay-Doppler plane to the time-frequency plane, where inter-symbol interference is removed, i.e., the user's channel matrix is diagonalized. The following proof confirms this intuition and reveals how the diagonalized channel matrix is related to the original block circulant matrix. We first apply $\mathbf{F}_N \otimes \mathbf{I}_M$ to $\mathbf{y}_0$, which yields the following:

$$\begin{aligned}
&(\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{y}_0 \\
&= (\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{H}_0\left(\gamma_0\mathbf{x}_0 + \sum_{q=1}^{M}\gamma_q\mathbf{x}_q\right) + (\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{z}_0 \\
&= \mathrm{diag}\left\{\sum_{n=0}^{N-1}\mathbf{A}_{0,n}e^{-j\frac{2\pi l n}{N}}, 0 \le l \le N-1\right\}(\mathbf{F}_N \otimes \mathbf{I}_M) \\
&\quad \times \left(\gamma_0\mathbf{x}_0 + \sum_{q=1}^{M}\gamma_q\mathbf{x}_q\right) + (\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{z}_0, \qquad (61)
\end{aligned}$$

where $\mathrm{diag}\{\mathbf{B}_1, \cdots, \mathbf{B}_N\}$ denotes a block-diagonal matrix with $\mathbf{B}_n$, $1 \le n \le N$, on its main diagonal. Note that $\sum_{n=0}^{N-1}\mathbf{A}_{0,n}e^{-j\frac{2\pi l n}{N}}$, $0 \le l \le N-1$, is a sum of N $M \times M$ circulant matrices, each of which can be further diagonalized by $\mathbf{F}_M$. Therefore, we can apply $\mathbf{I}_N \otimes \mathbf{F}_M^H$ to $(\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{y}_0$, which yields the following:

$$\begin{aligned}
&(\mathbf{I}_N \otimes \mathbf{F}_M^H)(\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{y}_0 \\
&\quad = \mathrm{diag}\left\{\sum_{n=0}^{N-1}\mathbf{\Lambda}_{0,n}e^{-j\frac{2\pi l n}{N}}, 0 \le l \le N-1\right\} \\
&\qquad \times (\mathbf{F}_N \otimes \mathbf{I}_M)(\mathbf{I}_N \otimes \mathbf{F}_M^H)\left(\gamma_0\mathbf{x}_0 + \sum_{q=1}^{M}\gamma_q\mathbf{x}_q\right) \\
&\qquad + (\mathbf{I}_N \otimes \mathbf{F}_M^H)(\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{z}_0, \qquad (62)
\end{aligned}$$

where $\mathbf{\Lambda}_{0,n}$ is a diagonal matrix, $\mathbf{\Lambda}_{0,n} = \mathrm{diag}\left\{\sum_{m=0}^{M-1}a_{0,n}^{m,1}e^{j\frac{2\pi t m}{M}}, 0 \le t \le M-1\right\}$, and $a_{0,n}^{m,1}$ is the element located in the m-th row and first column of $\mathbf{A}_{0,n}$.

By applying a property of the Kronecker product, $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$, the received signals can be simplified as follows:

$$\begin{aligned}
&(\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{y}_0 \\
&\quad = \underbrace{\mathrm{diag}\left\{\sum_{n=0}^{N-1}\mathbf{\Lambda}_{0,n}e^{-j\frac{2\pi l n}{N}}, 0 \le l \le N-1\right\}(\mathbf{F}_N \otimes \mathbf{F}_M^H)}_{\mathbf{D}_0} \\
&\qquad \times \left(\gamma_0\mathbf{x}_0 + \sum_{q=1}^{M}\gamma_q\mathbf{x}_q\right) + (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{z}_0, \qquad (63)
\end{aligned}$$

where the $(kM + l + 1)$-th element on the main diagonal of $\mathbf{D}_0$ is $D_0^{k,l}$ as defined in the proposition. The proof for the proposition is complete.

Appendix B
Proof for Lemma 1

In order to facilitate the SINR analysis, the system model in (18) is further simplified. Define $\tilde{X}[n,m] = \sum_{i=1}^{M} X_i[n,m]$. With the mapping scheme used in (6), the NOMA users' signals are interleaved and orthogonally placed in the time-frequency plane, i.e., $\tilde{X}[n,m]$ is simply U_{m+1}'s n-th signal, $x_{m+1}(n)$. Denote the outcome of the SFFT of $\tilde{X}[n,m]$ by $\tilde{x}[k,l]$, which yields the following transform:

$$\tilde{x}[k,l] = \frac{1}{\sqrt{NM}}\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}\tilde{X}[n,m]e^{-j2\pi\left(\frac{nk}{N}-\frac{ml}{M}\right)}. \qquad (64)$$

Denote the $NM \times 1$ vector collecting the $\tilde{x}[k,l]$ by $\tilde{\mathbf{x}}$ and the $NM \times 1$ vector collecting the $\tilde{X}[n,m]$ by $\breve{\mathbf{x}}$, which means that (64) can be rewritten as follows:

$$\tilde{\mathbf{x}} = (\mathbf{F}_N \otimes \mathbf{F}_M^H)\breve{\mathbf{x}}. \qquad (65)$$

Therefore, the model for the received signals in (18) can be re-written as follows:

$$\begin{aligned}
\breve{\mathbf{y}}_0 &= \gamma_0\mathbf{x}_0 + \gamma_1\tilde{\mathbf{x}} + \left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)^{-1}\mathbf{D}_0^{-1}\tilde{\mathbf{z}}_i \\
&= \gamma_0\mathbf{x}_0 + \underbrace{\gamma_1(\mathbf{F}_N \otimes \mathbf{F}_M^H)\breve{\mathbf{x}} + \left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)^{-1}\mathbf{D}_0^{-1}\tilde{\mathbf{z}}_0}_{\text{Interference and noise terms}}, \qquad (66)
\end{aligned}$$

where we have used the assumption that $\gamma_i = \gamma_1$, for $1 \le i \le N$. Note that the power of the information-bearing signals is simply $\gamma_0^2\rho$, and therefore, the key step to obtain the SINR is to find the covariance matrix of the interference-plus-noise term.

We first show that $\tilde{\mathbf{z}}_0 \triangleq (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{z}_0$ is still a complex Gaussian vector, i.e., $\tilde{\mathbf{z}}_i \sim CN(0, \mathbf{I}_{NM})$. Recall that $\mathbf{z}_0$ contains NM i.i.d. complex Gaussian random variables. Furthermore, $\mathbf{F}_N \otimes \mathbf{F}_M^H$ is a unitary matrix as shown in the following:

$$\begin{aligned}
(\mathbf{F}_N \otimes \mathbf{F}_M^H)(\mathbf{F}_N \otimes \mathbf{F}_M^H)^H &\overset{(a)}{=} (\mathbf{F}_N \otimes \mathbf{F}_M^H)(\mathbf{F}_N^H \otimes \mathbf{F}_M) \\
&\overset{(b)}{=} (\mathbf{F}_N\mathbf{F}_N^H) \otimes (\mathbf{F}_M^H\mathbf{F}_M) \\
&= \mathbf{I}_{NM}, \qquad (67)
\end{aligned}$$

where step (a) follows from the fact that $(\mathbf{A} \otimes \mathbf{B})^H = \mathbf{A}^H \otimes \mathbf{B}^H$ and step (b) follows from the fact that

$(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{AC})\otimes(\mathbf{BD})$. Therefore, $(\mathbf{F}_N\otimes\mathbf{F}_M^H)\mathbf{z}_0 \sim CN(0,\mathbf{I}_{NM})$ given the fact that $\mathbf{z}_0 \sim CN(0,\mathbf{I}_{NM})$ and a unitary transformation of a Gaussian vector is still a Gaussian vector.

Therefore, the covariance matrix of the interference-plus-noise term is given by

$$
\begin{aligned}
&\mathbf{C}_{\text{cov}} \\
&= \gamma_1^2\mathcal{E}\left\{(\mathbf{F}_N\otimes\mathbf{F}_M^H)\breve{\mathbf{x}}\breve{\mathbf{x}}^H\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)^H\right\} \\
&\quad+\mathcal{E}\left\{\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)^{-1}\mathbf{D}_0^{-1}\tilde{\mathbf{z}}_0\tilde{\mathbf{z}}_0^H\mathbf{D}_0^{-H}\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)^{-H}\right\}.
\end{aligned}
\tag{68}
$$

Recall that the $(nM+m+1)$-th element of $\breve{\mathbf{x}}$ is $\tilde{X}[n,m]$ which is equal to $x_{m+1}(n)$. Therefore, the covariance matrix can be further simplified as follows:

$$
\begin{aligned}
\mathbf{C}_{\text{cov}} &= \gamma_1^2\rho(\mathbf{F}_N\otimes\mathbf{F}_M^H)\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)^H \\
&\quad+\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)^{-1}\mathbf{D}_0^{-1}\mathbf{D}_0^{-H}\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)^{-H} \\
&= \gamma_1^2\rho\mathbf{I}_{MN}+\left(\mathbf{F}_N^H\otimes\mathbf{F}_M\right)\mathbf{D}_0^{-1}\mathbf{D}_0^{-H}\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right),
\end{aligned}
\tag{69}
$$

where the noise power is assumed to be normalized.

Following the same steps as in the proof of Proposition 1, we learn that, by construction, $\left(\mathbf{F}_N^H\otimes\mathbf{F}_M\right)\mathbf{D}_0^{-1}\mathbf{D}_0^{-H}\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)$ is also a block-circulant matrix, which means that the elements on the main diagonal of $\left(\mathbf{F}_N^H\otimes\mathbf{F}_M\right)\mathbf{D}_0^{-1}\mathbf{D}_0^{-H}\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)$ are identical. Without loss of generality, denote the diagonal elements of $\left(\mathbf{F}_N^H\otimes\mathbf{F}_M\right)\mathbf{D}_0^{-1}\mathbf{D}_0^{-H}\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)$ by ϕ. Therefore, ϕ can be found by using the trace of the matrix as follows:

$$
\begin{aligned}
\phi &= \frac{1}{NM}\mathrm{Tr}\left\{\left(\mathbf{F}_N^H\otimes\mathbf{F}_M\right)\mathbf{D}_0^{-1}\mathbf{D}_0^{-H}\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)\right\} \\
&= \frac{1}{NM}\mathrm{Tr}\left\{\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)\left(\mathbf{F}_N^H\otimes\mathbf{F}_M\right)\mathbf{D}_0^{-1}\mathbf{D}_0^{-H}\right\} \\
&= \frac{1}{NM}\mathrm{Tr}\left\{\mathbf{D}_0^{-1}\mathbf{D}_0^{-H}\right\} = \frac{1}{NM}\sum_{k=0}^{N-1}\sum_{l=0}^{M-1}|D_0^{k,l}|^{-2}.
\end{aligned}
\tag{70}
$$

Therefore, the SINR for detection of $x_0[k,l]$ is given by

$$
\mathrm{SINR}_{0,kl}^{LE} = \frac{\rho\gamma_0^2}{\rho\gamma_1^2+\phi},
\tag{71}
$$

and the proof is complete.

Appendix C
Proof for Lemma 2

The lemma is proved by first developing upper and lower bounds on the outage probability, and then showing that both bounds have the same diversity order.

An upper bound on $\mathrm{SINR}_{0,kl}$ is given by

$$
\begin{aligned}
\mathrm{SINR}_{0,kl} &= \frac{\rho\gamma_0^2}{\rho\gamma_1^2+\frac{1}{NM}\sum_{\tilde{k}=0}^{N-1}\sum_{\tilde{l}=0}^{M-1}|D_0^{\tilde{k},\tilde{l}}|^{-2}} \\
&\le \frac{\rho\gamma_0^2}{\rho\gamma_1^2+\frac{1}{NM}|D_0^{0,0}|^{-2}}.
\end{aligned}
\tag{72}
$$

Therefore, the outage probability, denoted by $\mathrm{P}_{0,kl}$, can be lower bounded as follows:

$$
\begin{aligned}
\mathrm{P}_{0,kl} &\ge \mathrm{P}\left(\frac{\rho\gamma_0^2}{\rho\gamma_1^2+\frac{1}{NM}|D_0^{0,0}|^{-2}}<\epsilon_0\right) \\
&= \mathrm{P}\left(|D_0^{0,0}|^2<\frac{\epsilon_0}{NM\rho(\gamma_0^2-\gamma_1^2\epsilon_0)}\right),
\end{aligned}
\tag{73}
$$

where we assume that $\gamma_0^2>\gamma_1^2\epsilon_0$. Otherwise, the outage probability is always one.

To evaluate the lower bound on the outage probability, the distribution of $D_0^{u,v}$ is required. Recall from (16) that $D_0^{u,v}$ is the $((v-1)M+u)$-th main diagonal element of $\mathbf{D}_0$ and can be expressed as follows:

$$
D_0^{u,v} = \sum_{n=0}^{N-1}\sum_{m=0}^{M-1}a_{0,n}^{m,1}e^{j2\pi\frac{um}{M}}e^{-j2\pi\frac{vn}{N}},
\tag{74}
$$

which is the ISFFT of $a_{0,n}^{m,1}$. Therefore, we have the following property:

$$
\tilde{\mathbf{D}}_0 = \sqrt{NM}\mathbf{F}_M^H\mathbf{A}_0\mathbf{F}_N,
\tag{75}
$$

where the element in the u-th row and the v-th column of $\tilde{\mathbf{D}}_0$ is $D_0^{u,v}$ and the element in the m-th row and the n-th column of $\mathbf{A}_0$ is $a_{0,n}^{m,1}$.

The matrix-based expression shown in (75) can be vectorized as follows:

$$
\begin{aligned}
\mathrm{Diag}(\mathbf{D}_0) &= \mathrm{vec}(\tilde{\mathbf{D}}_0) = \sqrt{NM}\mathrm{vec}(\mathbf{F}_M^H\mathbf{A}_0\mathbf{F}_N) \\
&= \sqrt{NM}(\mathbf{F}_N\otimes\mathbf{F}_M^H)\mathrm{vec}(\mathbf{A}_0),
\end{aligned}
\tag{76}
$$

where $\mathrm{Diag}(\mathbf{A})$ denotes a vector collecting all elements on the main diagonal of $\mathbf{A}$ and we use the facts that $(\mathbf{C}^T\otimes\mathbf{A})\mathrm{vec}(\mathbf{B}) = \mathrm{vec}(\mathbf{D})$ if $\mathbf{ABC}=\mathbf{D}$, and $\mathbf{F}_N^T=\mathbf{F}_N$.

We note that $\mathrm{vec}(\mathbf{A}_0)$ contains only (P_0+1) non-zero elements, where the remaining elements are zero. Therefore, each element on the main diagonal of $\mathbf{D}_0$ is a superposition of (P_0+1) i.i.d. random variables, $h_{i,p}\sim CN\left(0,\frac{1}{P_0+1}\right)$. We further note that the coefficients for the superposition are complex exponential constants, i.e., the magnitude of each coefficient is one. Therefore, each element on the main diagonal of $\mathbf{D}_0$ is still complex Gaussian distributed, i.e., $D_0^{u,v}\sim CN(0,1)$, which means that the lower bound on the outage probability shown in (73) can be expressed as follows:

$$
\mathrm{P}_{0,kl} \ge 1-e^{-\frac{\epsilon_0}{NM\rho(\gamma_0^2-\gamma_1^2\epsilon_0)}} \doteq \frac{1}{\rho}.
\tag{77}
$$

On the other hand, an upper bound on the outage probability is given by

$$
\mathrm{P}_{0,kl} \le \mathrm{P}\left(\frac{\rho\gamma_0^2}{\rho\gamma_1^2+\frac{1}{NM}\sum_{\tilde{k}=0}^{N-1}\sum_{\tilde{l}=0}^{M-1}|D_0^{\min}|^{-2}}<\epsilon_0\right),
\tag{78}
$$

where $|D_0^{\min}| = \min\{|D_0^{k,l}|,\forall l\in\{0,\cdots,M-1\},k\in\{0,\cdots,N-1\}\}$.

Therefore, the outage probability can be upper bounded as follows:

$$
\mathrm{P}_{0,kl} \le \mathrm{P}\left(|D_0^{\min}|^2<\frac{\epsilon_0}{\rho(\gamma_0^2-\gamma_1^2\epsilon_0)}\right).
\tag{79}
$$

It is important to point out that the $|D_0^{k,l}|^2$, $l \in \{0, \cdots, M-1\}, k \in \{0, \cdots, N-1\}$, are identically but not independently distributed. This correlation property is shown as follows. The covariance matrix of the effective channel gains, i.e., the elements on the main diagonal of $\mathbf{D}_0$, is given by

$$\begin{aligned}&\mathcal{E}\left\{\mathrm{Diag}(\mathbf{D}_0)\mathrm{Diag}(\mathbf{D}_0)^{\mathrm{H}}\right\}\\ &\quad = NM\mathcal{E}\left\{(\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathrm{vec}(\mathbf{A}_0)\mathrm{vec}(\mathbf{A}_0)^H(\mathbf{F}_N \otimes \mathbf{F}_M^H)^H\right\}\\ &\quad = NM(\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathcal{E}\left\{\mathrm{vec}(\mathbf{A}_0)\mathrm{vec}(\mathbf{A}_0)^H\right\}(\mathbf{F}_N \otimes \mathbf{F}_M^H)^H.\end{aligned} \tag{80}$$

Because the channel gains, $h_{0,p}$, are i.i.d., $\mathcal{E}\left\{\mathrm{vec}(\mathbf{A}_0)\mathrm{vec}(\mathbf{A}_0)^H\right\}$ is a diagonal matrix, where only (P_0+1) of its main diagonal elements are non-zero. Following the same steps as in the proof for Proposition 1, one can show that the product of $(\mathbf{F}_N \otimes \mathbf{F}_M^H)$, a diagonal matrix, and $(\mathbf{F}_N \otimes \mathbf{F}_M^H)^H$ yields a block circulant matrix, which means that $\mathcal{E}\left\{\mathrm{Diag}(\mathbf{D}_0)\mathrm{Diag}(\mathbf{D}_0)^{\mathrm{H}}\right\}$ is a block-circulant matrix, not a diagonal matrix. Therefore, the $|D_0^{k,l}|^2$, $l \in \{0, \cdots, M-1\}, k \in \{0, \cdots, N-1\}$, are correlated, and not independent.

Although the $|D_0^{k,l}|^2$ are not independent, an upper bound on $\mathrm{P}_{0,kl}$ can be still found as follows:

$$\begin{aligned}\mathrm{P}_{0,kl} &\leq \mathrm{P}\left(|D_0^{\min}|^2 < \frac{\epsilon_0}{\rho(\gamma_0^2 - \gamma_1^2\epsilon_0)}\right)\\ &\leq \sum_{k=0}^{N-1}\sum_{l=0}^{M-1}\mathrm{P}\left(|D_0^{k,l}|^2 < \frac{\epsilon_0}{\rho(\gamma_0^2 - \gamma_1^2\epsilon_0)}\right)\\ &\leq MN\mathrm{P}\left(|D_0^{0,0}|^2 < \frac{\epsilon_0}{\rho(\gamma_0^2 - \gamma_1^2\epsilon_0)}\right)\\ &= MN\left(1 - e^{-\frac{\epsilon_0}{\rho(\gamma_0^2 - \gamma_1^2\epsilon_0)}}\right) \doteq \frac{1}{\rho}.\end{aligned} \tag{81}$$

Since both the upper and lower bounds on the outage probability have the same diversity order, the proof of the lemma is complete.

References

[1] Z. Ding *et al.*, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[2] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 176–183, Oct. 2017.

[3] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[4] H. Sari, F. Vanhaverbeke, and M. Moeneclaey, "Multiple access using two sets of orthogonal signal waveforms," *IEEE Commun. Lett.*, vol. 4, no. 1, pp. 4–6, Jan. 2000.

[5] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE 24th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, London, U.K., Sep. 2013, pp. 611–615.

[6] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[7] 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.

[8] A. Brighente and S. Tomasin, "Power allocation for non-orthogonal millimeter wave systems with mixed traffic," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 432–443, Jan. 2019.

[9] Y. Wu, L. P. Qian, K. Ni, C. Zhang, and X. Shen, "Delay-minimization nonorthogonal multiple access enabled multi-user mobile edge computation offloading," *IEEE J. Sel. Topics Signal Process*, vol. 13, no. 3, pp. 392–407, Jun. 2019.

[10] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, 2017.

[11] Y. Zhou, V. W. S. Wong, and R. Schober, "Coverage and rate analysis of millimeter wave NOMA networks with beam misalignment," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8211–8227, Dec. 2018.

[12] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 629–633, May 2016.

[13] R. Chopra, C. R. Murthy, H. A. Suraweera, and E. G. Larsson, "Analysis of nonorthogonal training in massive MIMO under channel aging with SIC receivers," *IEEE Signal Process. Lett.*, vol. 26, no. 2, pp. 282–286, Feb. 2019.

[14] A. Maatouk, E. Çalışkan, M. Koca, M. Assaad, G. Gui, and H. Sari, "Frequency-domain NOMA with two sets of orthogonal signal waveforms," *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 906–909, May 2018.

[15] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[16] L. Yin, W. O. Popoola, X. Wu, and H. Haas, "Performance evaluation of non-orthogonal multiple access in visible light communication," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 5162–5175, Dec. 2016.

[17] H. Marshoud, V. M. Kapinas, G. K. Karagiannidis, and S. Muhaidat, "Non-orthogonal multiple access for visible light communications," *IEEE Photon. Technol. Lett.*, vol. 28, no. 1, pp. 51–54, Jan. 1, 2016.

[18] Z. Ding, P. Fan, and H. V. Poor, "Impact of non-orthogonal multiple access on the offloading of mobile edge computing," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 375–390, Jan. 2019.

[19] *Study on Downlink Multiuser Superposition Transmission for LTE*, document TR 36.859, 3GPP, Mar. 2015.

[20] *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation (Release 15)*, document TS 36.211, 3GPP, Jan. 2019.

[21] *Study on Non-Orthogonal Multiple Access (NOMA) for NR (Release 16)*, document TR 38.812, 3GPP, Dec. 2018.

[22] B. Di, L. Song, Y. Li, and Z. Han, "V2X meets NOMA: Non-orthogonal multiple access for 5G-enabled vehicular networks," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 14–21, Dec. 2017.

[23] Y. Chen, L. Wang, Y. Ai, B. Jiao, and L. Hanzo, "Performance analysis of NOMA-SM in vehicle-to-vehicle massive MIMO channels," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2653–2666, Dec. 2017.

[24] R. Hadani and A. Monk, "OTFS: A new generation of modulation addressing the challenges of 5G," 2018, *arXiv:1802.02623*. [Online]. Available: https://arxiv.org/abs/1802.02623

[25] R. Hadani *et al.*, "Orthogonal time frequency space modulation," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.

[26] K. R. Murali and A. Chockalingam, "On OTFS modulation for high-Doppler fading channels," in *Proc. Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, Feb. 2018, pp. 1–10.

[27] P. Raviteja, Y. Hong, E. Viterbo, and E. Biglieri, "Practical pulse-shaping waveforms for reduced-cyclic-prefix OTFS," *IEEE Trans. Veh. Tech.*, vol. 68, no. 1, pp. 957–961, Jan. 2019.

[28] P. Raviteja, K. T. Phan, Y. Hong, and E. Viterbo, "Interference cancellation and iterative detection for orthogonal time frequency space modulation," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6501–6515, Oct. 2018.

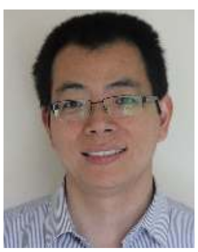
[29] G. D. Surabhi, R. M. Augustine, and A. Chockalingam, "On the diversity of uncoded OTFS modulation in doubly-dispersive channels," 2018, *arXiv:1808.07747*. [Online]. Available: https://arxiv.org/abs/1808.07747

[30] V. Khammammetti and S. K. Mohammed, "OTFS-based multiple-access in high Doppler and delay spread wireless channels," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 528–531, Apr. 2019.

[31] A. RezazadehReyhani, A. Farhang, M. Ji, R. R. Chen, and B. Farhang-Boroujeny, "Analysis of discrete-time MIMO OFDM-based orthogonal time frequency space modulation," in *Proc. IEEE Int. Conf. Communicat. (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–6.

[32] M. K. Ramachandran and A. Chockalingam, "MIMO-OTFS in high-Doppler fading channels: Signal detection and channel estimation," in *Proc. IEEE GLOBECOM*, Kansas City, MO, USA, Dec. 2018, pp. 206–212.
[33] P. Raviteja, E. Viterbo, and Y. Hong, "OTFS performance on static multipath channels," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 745–748, Jun. 2019.
[34] D. Falconer, S. L. Ariyavisitakul, A. Benyamin-Seeyar, and B. Eidson, "Frequency domain equalization for single-carrier broadband wireless systems," *IEEE Commun. Mag.*, vol. 40, no. 4, pp. 58–66, Apr. 2002.
[35] J. Louveaux, L. Vandendorpe, and T. Sartenaer, "Cyclic prefixed single carrier and multicarrier transmission: Bit rate comparison," *IEEE Commun. Lett.*, vol. 7, no. 4, pp. 180–182, Apr. 2003.
[36] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.
[37] B. Devillers, "Cyclic prefixed block transmission for wireless communications: Performance analysis and optimization," Ph.D. dissertation, Université catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium, 2009. AQ:5
[38] B. Devillers, J. Louveaux, and L. Vandendorpe, "About the diversity in cyclic prefixed single-carrier systems," *Phys. Commun.*, vol. 1, no. 4, pp. 266–276, Dec. 2008.
[39] J. Jeon, "NR wide bandwidth operations," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 42–46, Mar. 2018.
[40] C. Sexton, N. Marchetti, and L. A. DaSilva, "Customization and trade-offs in 5G RAN slicing," *IEEE Commun. Mag.*, vol. 57, no. 4, pp. 116–122, Apr. 2019.
[41] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "A new wireless communication paradigm through software-controlled metasurfaces," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 162–169, Sep. 2018.

AQ:6



Zhiguo Ding (S'03–M'05–SM'15) received the B.Eng. degree in electrical engineering from the Beijing University of Posts and Telecommunications in 2000 and the Ph.D. degree in electrical engineering from Imperial College London in 2005.

From July 2005 to April 2018, he was with Queen's University Belfast, Imperial College, Newcastle University, and Lancaster University. Since April 2018, he has been a Professor of communications with The University of Manchester. From October 2012 to September 2018, he was an Academic Visitor with Princeton University. His research interests are 5G networks, game theory, cooperative and energy harvesting networks, and statistical signal processing. He was a recipient of the Best Paper Award at IET ICWMC-2009 and IEEE WCSP-2014, the EU Marie Curie Fellowship (2012–2014), the Top IEEE TVT Editor 2017, the 2018 IEEE Communication Society Heinrich Hertz Award, the 2018 IEEE Vehicular Technology Society Jack Neubauer Memorial Award, and the 2018 IEEE Signal Processing Society Best Signal Processing Letter Award. He was an Editor of IEEE WIRELESS COMMUNICATIONS LETTERS and IEEE COMMUNICATIONS LETTERS from 2013 to 2016. He has been serving as an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and *Journal of Wireless Communications and Mobile Computing*.

Robert Schober (M'01–SM'08–F'10) received the Diploma (Univ.) and Ph.D. degrees in electrical engineering from the Friedrich-Alexander University of Erlangen-Nuremberg (FAU), Germany, in 1997 and 2000, respectively.

From 2002 to 2011, he was a Professor and the Canada Research Chair with The University of British Columbia (UBC), Vancouver, Canada. Since January 2012, he has been an Alexander von Humboldt Professor and the Chair for Digital Communication with FAU. His research interests fall into the broad areas of communication theory, wireless communications, and statistical signal processing. He is also a fellow of the Canadian Academy of Engineering and the Engineering Institute of Canada. He was a recipient of several awards for his work, including the 2002 Heinz Maier-Leibnitz Award of the German Science Foundation (DFG), the 2004 Innovations Award of the Vodafone Foundation for Research in Mobile Communications, the 2006 UBC Killam Research Prize, the 2007 Wilhelm Friedrich Bessel Research Award of the Alexander von Humboldt Foundation, the 2008 Charles McDowell Award for Excellence in Research from UBC, the 2011 Alexander von Humboldt Professorship, the 2012 NSERC E.W.R. Steacie Fellowship, and the 2017 Wireless Communications Recognition Award by the IEEE Wireless Communications Technical Committee. He was listed as a 2017 Highly Cited Researcher by the Web of Science. He is also the Chair of the Steering Committee of IEEE TRANSACTIONS ON MOLECULAR, BIOLOGICAL AND MULTI-SCALE COMMUNICATION, a member of the Editorial Board of PROCEEDINGS OF THE IEEE, a Member-at-Large of the Board of Governors of ComSoc, and the ComSoc Director of journals. He is also a Distinguished Lecturer of the IEEE Communications Society (ComSoc). From 2012 to 2015, he served as the Editor-in-Chief of IEEE TRANSACTIONS ON COMMUNICATIONS.

Pingzhi Fan (M'93–SM'99–F'15) received the M.Sc. degree in computer science from Southwest Jiaotong University, China, in 1987, and the Ph.D. degree in electronic engineering from Hull University, U.K., in 1994.

He was the Chief Scientist of the National 973 Research Project (MoST) from 2012 to 2016. He is currently a Professor and the Director of the Institute of Mobile Communications, Southwest Jiaotong University. He has been a Visiting Professor with Leeds University, U.K., since 1997, and has been a Guest Professor with Shanghai Jiaotong University since 1999. He has over 280 research papers published in various international journals and eight books (including edited). He is the inventor of 22 granted patents. His research interests include vehicular communications, wireless networks for big data, and signal design and coding. He is also a fellow of IET, CIE, and CIC. He was a recipient of the U.K. ORS Award in 1992 and the Outstanding Young Scientist Award (NSFC) in 1998. He has served as the general chair or the TPC chair of a number of international conferences. He is also the Founding Chair of IEEE VTS BJ Chapter, IEEE ComSoc CD Chapter, and IEEE Chengdu Section. He is also the guest editor or editorial member of several international journals. He has also served as the Board Member of IEEE Region 10, IET (IEE) Council, and IET Asia–Pacific Region. He is also an IEEE VTS Distinguished Lecturer (2015–2019).

H. Vincent Poor (M'77–SM'82–F'87) received the Ph.D. degree in EECS from Princeton University in 1977.

From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana–Champaign. Since 1990, he has been on the faculty at Princeton University, where he is currently the Michael Henry Strater University Professor of electrical engineering. From 2006 to 2016, he served as the Dean of the School of Engineering and Applied Science, Princeton University. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory and signal processing, and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the recent book *Multiple Access Techniques for 5G Wireless Networks and Beyond.* (Springer, 2019). He is also a member of the National Academy of Engineering and the National Academy of Sciences. He is also a Foreign Member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He was a recipient of the Marconi and Armstrong Awards of the IEEE Communications Society in 2007 and 2009, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal, the 2019 ASEE Benjamin Garver Lamme Award, the D.Sc. *(honoris causa)* from Syracuse University awarded in 2017, and the D.Eng. *(honoris causa)* from the University of Waterloo awarded in 2019.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your article. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

AQ:1 = Please supply index terms/keywords for your paper. To download the IEEE Taxonomy, go to http://www.ieee.org/documents/taxonomy_v101.pdf.

AQ:2 = Author: Please confirm or add details for any funding or financial support for the research of this article.

AQ:3 = Please provide the postal code for The University of Manchester, Friedrich-Alexander-University Erlangen-Nurnberg, and Southwest Jiaotong University.

AQ:4 = Note that if you require corrections/changes to tables or figures, you must supply the revised files, as these items are not edited for you.

AQ:5 = Please provide the department name for Ref. [37].

AQ:6 = Current affiliation in the biography of Zhiguo Ding does not match the First Footnote. Please check and correct where needed.

OTFS-NOMA: An Efficient Approach for Exploiting Heterogenous User Mobility Profiles

Zhiguo Ding, *Senior Member, IEEE*, Robert Schober, *Fellow, IEEE*,
Pingzhi Fan, *Fellow, IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*— This paper considers a challenging communication scenario, in which users have heterogenous mobility profiles, e.g., some users are moving at high speeds and some users are static. A new non-orthogonal multiple-access (NOMA) transmission protocol that incorporates orthogonal time frequency space (OTFS) modulation is proposed. Thereby, users with different mobility profiles are grouped together for the implementation of NOMA. The proposed OTFS-NOMA protocol is shown to be applicable to both uplink and downlink transmission, where sophisticated transmit and receive strategies are developed to remove inter-symbol interference and harvest both multi-path and multi-user diversity. Analytical results demonstrate that both the high-mobility and the low-mobility users benefit from the application of OTFS-NOMA. In particular, the use of NOMA allows the spreading of the high-mobility users' signals over a large amount of time-frequency resources, which enhances the OTFS resolution and improves the detection reliability. In addition, OTFS-NOMA ensures that low-mobility users have access to bandwidth resources which in conventional OTFS-orthogonal multiple access (OTFS-OMA) would be solely occupied by the high-mobility users. Thus, OTFS-NOMA improves the spectral efficiency and reduces latency.**

AQ:1 ***Index Terms*— XXXXX.**

I. Introduction

NON-ORTHOGONAL multiple access (NOMA) has been recognized as a paradigm shift for the design of multiple access techniques for the next generation of wireless networks [1]–[4]. Many existing works on NOMA have focused on scenarios with low-mobility users, where users with different channel conditions or quality of service (QoS) requirements are grouped together for the implementation of NOMA. For example, in power-domain NOMA, a base station serves two users simultaneously [5], [6]. In particular, the base station first orders the users according to their channel conditions, where the 'weak user' which has a poorer connection to the base station is generally allocated more transmission power and the other user, referred to as the 'strong user', is allocated less power. As such, the two users can be served in the same time-frequency resource, which improves the spectral efficiency compared to orthogonal multiple access (OMA). In the case that users have similar channel conditions, grouping users with different QoS requirements can facilitate the implementation of NOMA and effectively exploit the potential of NOMA [7]–[9]. Various existing studies have shown that the NOMA principle can be applied to different communication networks, such as millimeter-wave networks [10], [11], massive multiple-input multiple-output (MIMO) systems [12], [13], hybrid multiple access systems [14], [15], visible light communication networks [16], [17], and mobile edge computing [18]. We also note that various standardization efforts have been made to facilitate the implementation of NOMA in practical systems. For example, a study for the application of NOMA for downlink transmission, termed multi-user superposition transmission (MUST), was carried out for the 3rd Generation Partnership Project (3GPP) Release 14, where 15 different forms of MUST were proposed and compared [19]. After this study was completed, MUST was formally included in 3GPP Release 15 which is also referred to as Evolved Universal Terrestrial Radio Access (E-UTRA) [20]. A study for the application of NOMA for uplink transmission has been recently carried out for 3GPP Release 16, where more than 20 different forms of NOMA have been proposed by various companies [21].

This paper considers the application of NOMA to a challenging communication scenario, where users have heterogeneous mobility profiles. Different from the existing works in [22], [23], the use of orthogonal time frequency space (OTFS) modulation is considered in this paper because of its superior performance in scenarios with doubly-dispersive channels [24]–[26]. Recall that the key idea of OTFS is to use the delay-Doppler plane, where the users' signals are

Manuscript received April 4, 2019; revised June 5, 2019; accepted
AQ:2 July 14, 2019. The work of Z. Ding was supported by the UK Engineering and Physical Sciences Research Council under grant number EP/P009719/2 and by H2020-MSCA-RISE-2015 under grant number 690750. The work of P. Fan was supported by the National Natural Science Foundation of China under grant number 61731017, and the 111 Project (No.111-2-14). The work of H. V. Poor was supported by the U.S. National Science Foundation under Grants CCF-093970 and CCF-1513915. The associate editor coordinating the review of this article and approving it for publication was V. Raghavan. *(Corresponding author: Zhiguo Ding.)*

Z. Ding is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA, and also with the School of Electrical
AQ:3 and Electronic Engineering, The University of Manchester, Manchester, U.K. (e-mail: zhiguo.ding@manchester.ac.uk).

R. Schober is with the Institute for Digital Communications, Friedrich-Alexander-University Erlangen-Nurnberg (FAU), Erlangen, Germany (e-mail: robert.schober@fau.de).

P. Fan is with the Institute of Mobile Communications, Southwest Jiaotong University, Chengdu, China (e-mail: pingzhifan@foxmail.com).

H. V. Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Digital Object Identifier 10.1109/TCOMM.2019.2932934

orthogonally placed. Compared to conventional modulation schemes, such as orthogonal frequency-division multiplexing (OFDM), OTFS offers the benefit that the time-invariant channel gains in the delay-Doppler plane can be utilized, which simplifies channel estimation and signal detection in high-mobility scenarios. The impact of pulse-shaping waveforms on the performance of OTFS was studied in [27], and the design of interference cancellation and iterative detection for OTFS was investigated in [28]. The diversity gain achieved by OTFS was studied in [29], and the application of OTFS to multiple access was proposed in [30]. In [31] and [32], the concept of OTFS was combined with MIMO, which revealed that the use of spatial degrees of freedom can further enhance the performance of OTFS.

This paper considers the application of OTFS to NOMA communication networks, where the coexistence of NOMA and OTFS is investigated. In particular, this paper makes the following contributions:

1) A spectrally efficient OTFS-NOMA transmission protocol is proposed by grouping users with different mobility profiles for the implementation of NOMA. On the one hand, users with high mobility are served in the delay-Doppler plane, and their signals are modulated by OTFS. On the other hand, users with low mobility are served in the time-frequency plane, and their signals are modulated in a manner similar to conventional OFDM.
2) The proposed new OTFS-NOMA protocol is applied to both uplink and downlink transmission, where different rate and power allocation policies are used to suppress multiple access interference. In addition, sophisticated equalization techniques, such as the frequency-domain zero-forcing linear equalizer (FD-LE) and the decision feedback equalizer (FD-DFE), are employed to remove the inter-symbol interference in the delay-Doppler plane. The impact of the developed equalization techniques on OTFS-NOMA is analyzed by using the outage probability as the performance criterion. Strategies to harvest multi-path diversity and multi-user diversity are also introduced, which can further improve the outage performance of OTFS-NOMA transmission.
3) The developed analytical results demonstrate that both the high-mobility and the low-mobility users benefit from the proposed OTFS-NOMA scheme. The use of NOMA allows the high-mobility users' signals to be spread over a large amount of time-frequency resources without degrading the spectral efficiency. As a result, the OTFS resolution, which determines whether the users' channels can be accurately located in the delay-Doppler plane, is enhanced significantly, and therefore, the reliability of detecting the high-mobility users' signals is improved. We note that, in OTFS-OMA, enhancing the OTFS resolution implies that a large amount of time and frequency resources are solely occupied by the high-mobility users, which reduces the overall spectral efficiency since the high-mobility users' channel conditions are typically weaker than those of the low-mobility users. In contrast, the use of OTFS-NOMA ensures that the low-mobility users can access the bandwidth resources which would be solely occupied by the high-mobility users in the OMA mode. Hence, OTFS-NOMA improves spectral efficiency and reduces latency, as with OTFS-OMA the low-mobility users may have to wait for a long time before the scarce bandwidth resources occupied by the high-mobility users become available. In addition, we note that for the low-mobility users, using OFDM yields the same reception reliability as using OTFS, as pointed out in [33]. Therefore, the proposed OTFS-NOMA scheme, which serves the low-mobility users in the time-frequency plane and modulates the low-mobility users' signals in a manner similar to OFDM, offers the same reception reliability as OTFS-OMA, which serves the low-mobility users in the delay-Doppler plane and modulates the low-mobility users' signals by OTFS. However, OTFS-NOMA has the benefit of reduced system complexity because the use of the complicated OTFS transforms is avoided.

II. Foundations of OTFS-NOMA

A. Time-Frequency Plane and Delay-Doppler Plane

The key idea of OTFS-NOMA is to efficiently use both the time-frequency plane and the delay-Doppler plane. A discrete time-frequency plane is obtained by sampling at intervals of T s and Δf Hz as follows:

$$\Lambda_{\mathrm{TF}} = \{(nT, m\Delta f), n=0,\cdots,N-1, m=0,\cdots,M-1\}, \tag{1}$$

and the corresponding discrete delay-Doppler plane is given by

$$\Lambda_{\mathrm{DD}} = \left\{\left(\frac{k}{NT}, \frac{l}{M\Delta f}\right), k=0,\cdots,N-1, l=0,\cdots,M-1\right\}, \tag{2}$$

where N and M denote the total number of time intervals and the total number of frequency subchannels, respectively. The choices for T and Δf are determined by the channel characteristics, as will be explained in the following subsection.

B. Channel Model

This paper considers a multi-user communication network in which one base station communicates with $(K+1)$ users, denoted by U_i, $0 \leq i \leq K$. Denote U_i's channel response in the delay-Doppler plane by $h_i(\tau,\nu)$, where τ denotes the delay and ν denotes the Doppler shift. OTFS uses the sparsity feature of a wireless channel in the delay-Doppler plane, i.e., there are a small number of propagation paths between a transmitter and a receiver [24], [25], [28], which means that $h_i(\tau,\nu)$ can be expressed as follows:

$$h_i(\tau,\nu) = \sum_{p=0}^{P_i} h_{i,p}\delta(\tau-\tau_{i,p})\delta(\nu-\nu_{i,p}), \tag{3}$$

where (P_i+1) denotes the number of propagation paths, and $h_{i,p}$, $\tau_{i,p}$, and $\nu_{i,p}$ denote the complex Gaussian channel gain,[1] the delay, and the Doppler shift associated with the

[1]The Gaussian assumption has been commonly used in the OTFS literature [26]–[29] since each channel gain (or each tap of the delay-Doppler impulse response) represents a cluster of reflectors with specific delay and Doppler characteristics.

p-th propagation path, respectively. We assume that the $h_{i,p}$, $0 \leq p \leq P_i$, are independent and identically distributed (i.i.d.) random variables,[2] i.e., $h_{i,p} \sim CN\left(0, \frac{1}{P_i+1}\right)$, which means $\sum_{p=0}^{P_i} \mathcal{E}\{|h_{i,p}|^2\} = 1$, where $\mathcal{E}\{\cdot\}$ denotes the expectation operation. The discrete delay and Doppler tap indices for the p-th path of $h_i(\tau, \nu)$, denoted by $l_{\tau_{i,p}}$ and $k_{\nu_{i,p}}$, respectively, are given by [28]

$$\tau_{i,p} = \frac{l_{\tau_{i,p}} + \tilde{l}_{\tau_{i,p}}}{M\Delta f}, \quad \nu_{i,p} = \frac{k_{\nu_{i,p}} + \tilde{k}_{\nu_{i,p}}}{NT}, \tag{4}$$

where $\tilde{l}_{\tau_{i,p}}$ and $\tilde{k}_{\nu_{i,p}}$ denote the fractional delay and the fractional Doppler shift, respectively.

The construction of Λ_{TF} and Λ_{DD} needs to ensure that T is not smaller than the maximal delay spread, and Δf is not smaller than the largest Doppler shift, i.e., $T \geq \max\{\tau_{i,p}, 0 \leq p \leq P_i, 0 \leq i \leq K\}$ and $\Delta f \geq \max\{\nu_{i,p}, 0 \leq p \leq P_i, 0 \leq i \leq K\}$. In addition, the choices of N and M affect the OTFS resolution, which determines whether $h_i(\tau, \nu)$ can be accurately located in the discrete delay-Doppler plane. In particular, M and N need to be sufficiently large to approximately achieve ideal OTFS resolution, which ensures that $\tilde{l}_{\tau_{i,p}} = \tilde{k}_{\nu_{i,p}} = 0$, such that the interference caused by fractional delay and Doppler shift is effectively suppressed [24].

C. General Principle of OTFS-NOMA

To facilitate the illustration of the general principle of OTFS-NOMA, we first briefly describe OTFS-OMA, the benchmark scheme used in this paper. In OTFS-OMA, there is no spectrum sharing between the high-mobility users and the low-mobility users, i.e., if OTFS is used to serve the high-mobility users, the NT time intervals and the $M\Delta f$ frequency subchannels are occupied by the high-mobility users and the low-mobility users cannot be served in these resource blocks. The general principle of the proposed OTFS-NOMA scheme is to exploit both the delay-Doppler plane and the time-frequency plane, where users with heterogenous mobility profiles are grouped together and served simultaneously. On the one hand, for the users with high mobility, their signals are placed in the delay-Doppler plane, which means that the time-invariant channel gains in the delay-Doppler plane can be exploited. It is worth pointing out that in order to ensure that the channels can be located in the delay-Doppler plane, both N and M need to be large, which is a disadvantage of OTFS-OMA, since a significant number of frequency channels (e.g., $M\Delta f$) are occupied for a long time (e.g., NT) by the high-mobility users whose channel conditions can be quite weak. The use of OTFS-NOMA facilitates spectrum sharing and hence ensures that the high-mobility users' signals can be spread over a large amount of time-frequency resources without degrading the spectral efficiency.

On the other hand, for the users with low mobility, their signals are placed in the time-frequency plane. The interference between the users with different mobility profiles is managed by using the principle of NOMA. As a result, compared to OTFS-OMA, OTFS-NOMA improves the overall spectral efficiency since it encourages spectrum sharing among users with different mobility profiles and avoids that the bandwidth resources are solely occupied by the high-mobility users which might have weak channel conditions. In addition, the complexity of detecting the low-mobility users' signals is reduced, compared to OTFS-OMA which serves all users in the delay-Doppler plane.

In this paper, we assume that, among $(K+1)$ users, U_0 is a user with high mobility, and the remaining K users, U_i for $1 \leq i \leq K$, are low-mobility users, which are referred to as 'NOMA' users.[3] For OTFS-OMA, we assume that U_0 solely occupies all NM resource blocks in Λ_{DD}. In OTFS-NOMA, U_i, for $1 \leq i \leq K$, are opportunistic NOMA users and their signals are placed in Λ_{TF}. The design of downlink OTFS-NOMA transmission will be discussed in detail in Sections III, IV, and V. The application of OTFS-NOMA for uplink transmission will be considered in Section VI only briefly, due to space limitations.

III. Downlink OFTS-NOMA - System Model

In this section, the OTFS-NOMA downlink transmission protocol is described. In particular, assume that the base station sends NM symbols to U_0, denoted by $x_0[k,l]$, $k \in \{0, \cdots, N-1\}$, $l \in \{0, \cdots, M-1\}$. By using the inverse symplectic finite Fourier transform (ISFFT), the high-mobility user's symbols placed in the delay-Doppler plane are converted to NM symbols in the time-frequency plane as follows [24]:

$$X_0[n,m] = \frac{1}{NM} \sum_{k=0}^{N-1} \sum_{l=0}^{M-1} x_0[k,l] e^{j2\pi\left(\frac{kn}{N} - \frac{ml}{M}\right)}, \tag{5}$$

where $n \in \{0, \cdots, N-1\}$ and $m \in \{0, \cdots, M-1\}$. We note that the NM time-frequency signals can be viewed as N OFDM symbols containing M signals each. We assume that a rectangular window is applied to the transmitted and received signals.

The NOMA users' signals are placed directly in the time-frequency plane, and are superimposed with the high-mobility user's signals, $X_0[n,m]$. With NM orthogonal resource blocks available in the time-frequency plane, there are different ways for the K users to share the resource blocks. For illustration purposes, we assume that M users are selected from the K opportunistic NOMA users,[4] where each NOMA

[2] In order to simplify the performance analysis, we assume that the users' channels are i.i.d. In practice, it is likely that the high-mobility users' channel conditions are worse than the low-mobility users' channel conditions. This channel difference is beneficial for the implementation of NOMA, and hence can further increase the performance gain of OTFS-NOMA over OTFS-OMA.

[3] We note that the principle of OTFS-NOMA can be extended to the case where multiple high-mobility users are served in the delay-Doppler plane. In this case, the NM signals in the delay-Doppler plane belong to different high-mobility users and OTFS is used as a type of multiple access technique [24], [30]. For downlink transmission, this change has no impact on the proposed detection schemes and the analytical results developed in this paper. For uplink transmission, the results developed in this paper are applicable to the case with multiple high-mobility users if the adaptive-rate transmission scheme proposed in Section VI is employed.

[4] The same M users can be scheduled as long as the users' channels do not change in the delay-Doppler plane. Otherwise, a new set of M users may be selected from the K opportunistic users. We also note that the number of the opportunistic users is assumed to be larger than the number of the frequency subchannels ($K \geq M$), which can be justified by a spectrum crunch scenario, i.e., there are not sufficient bandwidth resources available to support a large number of mobile devices.

user is to occupy one frequency subchannel and receive N information bearing symbols, denoted by $x_i(n)$, for $1 \le i \le M$ and $0 \le n \le N-1$. The criterion employed for user scheduling and its impact on the performance of OTFS-NOMA will be discussed in Section V. Denote the time-frequency signals to be sent to U_i by $X_i[n,m]$, $1 \le i \le M$. The following mapping scheme is used in this paper[5]:

$$X_i[n,m] = \begin{cases} x_i(n) & \text{if} \quad m = i-1 \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

for $1 \le i \le M$ and $0 \le n \le N-1$.

The base station superimposes U_0's time-frequency signals with the NOMA users' signals as follows:

$$X[n,m] = \frac{\gamma_0}{NM}\sum_{k=0}^{N-1}\sum_{l=0}^{M-1} x_0[k,l]e^{j2\pi\left(\frac{kn}{N}-\frac{ml}{M}\right)} + \sum_{i=1}^{M}\gamma_i X_i[n,m], \tag{7}$$

where γ_i denotes the NOMA power allocation coefficient of user i, and $\sum_{i=0}^{M}\gamma_i^2 = 1$.

The transmitted signal at the base station is obtained by applying the Heisenberg transform to $X[n,m]$. By assuming perfect orthogonality between the transmit and receive pulses, the received signal at U_i in the time-frequency plane can be modelled as follows [24], [25], [28]:

$$Y_i[n,m] = H_i[n,m]X[n,m] + W_i[n,m], \tag{8}$$

where $W_i(n,m)$ is the white Gaussian noise in the time-frequency plane, and $H_i(n,m) = \int\int h_i(\tau,\nu)e^{j2\pi\nu nT} e^{-j2\pi(\nu+m\Delta f)\tau}d\tau d\nu$.

IV. Downlink OTFS-NOMA - Detecting the High-Mobility User's Signals

For the proposed downlink OTFS-NOMA scheme, U_0 directly detects its signals in the delay-Doppler plane by treating the NOMA users' signals as noise. In particular, in order to detect U_0's signals, the symplectic finite Fourier transform (SFFT) is applied to $Y_0[n,m]$ to obtain the delay-Doppler estimates as follows:

$$\begin{aligned} y_0[k,l] &= \frac{1}{NM}\sum_{n=0}^{N-1}\sum_{m=0}^{M-1} Y_0[n,m]e^{-j2\pi\left(\frac{nk}{N}-\frac{ml}{M}\right)} \\ &= \frac{1}{NM}\sum_{q=0}^{M}\gamma_q\sum_{n=0}^{N-1}\sum_{m=0}^{M-1} x_q[n,m]h_{w,0}\left(\frac{k-n}{NT},\frac{l-m}{M\Delta f}\right) \\ &\quad + z_0[k,l], \end{aligned} \tag{9}$$

where q denotes the user index, $z_0[k,l]$ is complex Gaussian noise, $x_q[k,l]$, $1 \le q \le M$, denotes the delay-Doppler representation of $X_q[n,m]$ and can be obtained by applying the SFFT to $X_q[n,m]$, the channel $h_{w,0}(\nu',\tau')$ is given by

$$h_{w,0}(\nu',\tau') = \int\int h_i(\tau,\nu)w(\nu'-\nu,\tau'-\tau)e^{-j2\pi\nu\tau}d\tau d\nu, \tag{10}$$

and $w(\nu,\tau) = \sum_{c=0}^{N-1}\sum_{d=0}^{M-1}e^{-j2\pi(\nu cT-\tau d\Delta f)}$. To simplify the analysis, the power of the complex-Gaussian distributed noise is assumed to be normalized, i.e., $z_i[k,l] \sim CN(0,1)$, where $CN(a,b)$ denotes a complex Gaussian distributed random variable with mean a and variance b.

By applying the channel model in (3), the relationship between the transmitted signals and the observations in the delay-Doppler plane can be expressed as follows [24], [25], [28]:

$$y_0[k,l] = \sum_{q=0}^{M}\gamma_q\sum_{p=0}^{P_0} h_{0,p}x_q[(k-k_{\nu_{0,p}})_N,(l-l_{\tau_{0,p}})_M] + z_0[k,l], \tag{11}$$

where $(\cdot)_N$ denotes the modulo N operator. As in [29]–[31], we assume that N and M are sufficiently large to ensure that both $\tilde{k}_{\nu_{0,p}}$ and $\tilde{l}_{\tau_{0,p}}$ are zero, i.e., there is no interference caused by fractional delay or fractional Doppler shift. We note that for OTFS-OMA, increasing N and M can significantly reduce spectral efficiency, whereas the use of large N and M becomes possible for OTFS-NOMA because of the spectrum sharing of users with different mobility profiles.

Define $\mathbf{y}_{0,k} = \left[y_0[k,0]\cdots y_0[k,M-1]\right]^T$ and $\mathbf{y}_0 = \left[\mathbf{y}_{0,0}^T\cdots\mathbf{y}_{0,N-1}^T\right]^T$. Similarly, the signal vector $\mathbf{x}_i$ and the noise vector $\mathbf{z}_0$ are constructed from $x_i[k,l]$ and $z_0[k,l]$, respectively. Based on (11), the system model can be expressed in matrix form as follows:

$$\mathbf{y}_0 = \gamma_0\mathbf{H}_0\mathbf{x}_0 + \underbrace{\sum_{q=1}^{M}\gamma_q\mathbf{H}_0\mathbf{x}_q + \mathbf{z}_0}_{\text{Interference and noise terms}}, \tag{12}$$

where $\mathbf{H}_0$ is a block-circulant matrix and defined as follows:

$$\mathbf{H}_0 = \begin{bmatrix} \mathbf{A}_{0,0} & \mathbf{A}_{0,N-1} & \cdots & \mathbf{A}_{0,2} & \mathbf{A}_{0,1} \\ \mathbf{A}_{0,1} & \mathbf{A}_{0,0} & \ddots & \mathbf{A}_{0,3} & \mathbf{A}_{0,2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{A}_{0,N-2} & \mathbf{A}_{0,N-3} & \ddots & \mathbf{A}_{0,0} & \mathbf{A}_{0,N-1} \\ \mathbf{A}_{0,N-1} & \mathbf{A}_{0,N-2} & \ddots & \mathbf{A}_{0,1} & \mathbf{A}_{0,0} \end{bmatrix}, \tag{13}$$

and each submatrix $\mathbf{A}_{0,n}$ is an $M\times M$ circulant matrix whose structure is determined by (11).

Example: Consider a special case with $N = 4$ and $M = 3$, and U_0's channel is given by

$$h_0(\tau,\nu) = h_{0,0}\delta(\tau)\delta(\nu) + h_{0,1}\delta\left(\tau-\frac{1}{M\Delta f}\right)\delta\left(\nu-\frac{3}{NT}\right), \tag{14}$$

which means $k_0 = 0$, $k_1 = 3$, $l_0 = 0$, $l_1 = 1$. Therefore, the block-circulant matrix is given by

$$\mathbf{H}_0 = \begin{bmatrix} \mathbf{A}_{0,0} & \mathbf{A}_{0,3} & \mathbf{A}_{0,2} & \mathbf{A}_{0,1} \\ \mathbf{A}_{0,1} & \mathbf{A}_{0,0} & \mathbf{A}_{0,3} & \mathbf{A}_{0,2} \\ \mathbf{A}_{0,2} & \mathbf{A}_{0,1} & \mathbf{A}_{0,0} & \mathbf{A}_{0,3} \\ \mathbf{A}_{0,3} & \mathbf{A}_{0,2} & \mathbf{A}_{0,1} & \mathbf{A}_{0,0} \end{bmatrix}, \tag{15}$$

[5] We note that mapping schemes different from (6) can also be used. For example, if N users are scheduled and each user is to occupy one time slot and receives an OFDM-like symbol containing M signals, we can set $X_i[n,m] = x_i(m)$, for $n = i-1$.

where $\mathbf{A}_{0,0} = h_{0,0}\mathbf{I}_3$, $\mathbf{A}_{0,1} = \mathbf{A}_{0,2} = \mathbf{0}_{3\times 3}$ and $\mathbf{A}_{0,3} = \begin{bmatrix} 0 & 0 & h_{0,1} \\ h_{0,1} & 0 & 0 \\ 0 & h_{0,1} & 0 \end{bmatrix}$.

Remark 1: It is well known that an $n\times n$ circulant matrix can be diagonalized by the $n \times n$ discrete Fourier transform (DFT) and inverse DFT matrices, denoted by $\mathbf{F}_n$ and $\mathbf{F}_n^{-1}$, respectively, i.e., the columns of the DFT matrix are the eigenvectors of the circulant matrix. We note that directly applying the DFT factorization to $\mathbf{H}_0$ is not possible, since $\mathbf{H}_0$ is not a circulant matrix, but a block circulant matrix.

Because of the structure of $\mathbf{H}_0$, inter-symbol interference still exists in the considered OTFS-NOMA system, and equalization is needed. We consider two equalization approaches, FD-LE and FD-DFE, which were both originally developed for single-carrier transmission with cyclic prefix [34], [35].

A. Design and Performance of FD-LE

The proposed FD-LE consists of two steps. Let $\otimes$ denote the Kronecker product. The first step is to multiply the observation vector $\mathbf{y}_0$ by $\mathbf{F}_N \otimes \mathbf{F}_M^H$, which leads to the result in the following proposition.

Proposition 1: By applying the detection matrix $\mathbf{F}_N \otimes \mathbf{F}_M^H$ to observation vector $\mathbf{y}_0$, the received signals for OTFS-NOMA downlink transmission can be written as follows:

$$\tilde{\mathbf{y}}_0 = \mathbf{D}_0(\mathbf{F}_N \otimes \mathbf{F}_M^H)\left(\gamma_0\mathbf{x}_0 + \sum_{q=1}^{M}\gamma_q\mathbf{x}_q\right) + \tilde{\mathbf{z}}_0, \quad (16)$$

where $\tilde{\mathbf{y}}_0 = (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{y}_0$, $\tilde{\mathbf{z}}_0 = (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{z}_0$, $\mathbf{D}_0$ is a diagonal matrix whose $(kM+l+1)$-th main diagonal element is given by

$$D_0^{k,l} = \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} a_{0,n}^{m,1} e^{j2\pi\frac{lm}{M}} e^{-j2\pi\frac{kn}{N}}, \quad (17)$$

for $0 \leq k \leq N-1$, $0 \leq l \leq M-1$, and $a_{0,n}^{m,1}$ is the element located in the $(nM+m+1)$-th row and the first column of $\mathbf{H}_0$.

Proof: Please refer to Appendix A. □

With the simplified signal model shown in (16), the second step of FD-LE is to apply $\left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)^{-1}\mathbf{D}_0^{-1}$ to $\tilde{\mathbf{y}}_0$. Thus, U_0's received signal is given by

$$\check{\mathbf{y}}_0 = \gamma_0\mathbf{x}_0 + \underbrace{\sum_{q=1}^{M}\gamma_q\mathbf{x}_q + \left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)^{-1}\mathbf{D}_0^{-1}\tilde{\mathbf{z}}_0}_{\text{Interference and noise terms}}, \quad (18)$$

where $\check{\mathbf{y}}_0 = \left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)^{-1}\mathbf{D}_0^{-1}\tilde{\mathbf{y}}_0$. To simplify the analysis, we assume that the powers of all users' information-bearing signals are identical, which means that the transmit signal-to-noise ratio (SNR) can be defined as $\rho = \mathcal{E}\{|x_0[k,l]|^2\} = \mathcal{E}\{|x_i(n)|^2\}$, since the noise power is assumed to be normalized.[6] The following lemma provides the signal-to-interference-plus-noise ratio (SINR) achieved by FD-LE.

[6]Following steps in the proof for Proposition 1 to show the statistical property of $\tilde{\mathbf{z}}_0$ in (66), we can also show that $W_i[n,m] \sim CN(0,1)$ if $z_i[k,l] \sim CN(0,1)$.

Lemma 1: Assume that $\gamma_i = \gamma_1$, for $1 \leq i \leq N$. By using FD-LE, the SINRs for detecting all $x_0[k,l]$, $0 \leq k \leq N-1$ and $0 \leq l \leq M-1$, are identical and given by

$$\mathit{SINR}_{0,kl}^{LE} = \frac{\rho\gamma_0^2}{\rho\gamma_1^2 + \frac{1}{NM}\sum_{\tilde{k}=0}^{N-1}\sum_{\tilde{l}=0}^{M-1}|D_0^{\tilde{k},\tilde{l}}|^{-2}}. \quad (19)$$

Proof: Please refer to Appendix B. □

Remark 2: The proof of Lemma 1 shows that $\sum_{i=0}^{M}\gamma_i^2 = 1$ can be simplified as $\gamma_0^2 + \gamma_i^2 = 1$ for $1 \leq i \leq M$, which is the motivation for assuming $\gamma_i = \gamma_1$. Following steps similar to those in the proofs for Proposition 1 and Lemma 1, one can show that directly applying $\mathbf{H}_0^{-1}$ to the observation vector yields the same SINR. However, the proposed FD-LE scheme can be implemented more efficiently since $\left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)^{-1} = \mathbf{F}_N^H \otimes \mathbf{F}_M$ and $\mathbf{D}_0$ is a diagonal matrix. Hence, the inversion of a full $NM \times NM$ matrix is avoided.

In this paper, the outage probability and the outage rate are used as performance criteria, since the outage probability can provide a tight bound on the probability of erroneous detection and is general in the sense that it does not depend on particular channel coding and modulation schemes used [36]. The outage probability achieved by FD-LE is given by $\mathrm{P}(\log(1 + \mathrm{SINR}_{0,kl}^{\mathrm{LE}}) < R_0)$, where R_i, $0 \leq i \leq M$, denotes U_i's target data rate. It is difficult to analyze the outage probability for the following two reasons. First, the $D_0^{k,l}$, $k \in \{0, \cdots, N-1\}$, $l \in \{0, \cdots, M-1\}$, are not statistically independent, and second, the distribution of a sum of the inverse of exponentially distributed random variables is difficult to characterize. The following lemma provides an asymptotic result for the outage probability based on the SINR provided in Lemma 1.

Lemma 2: If $\gamma_0^2 > \gamma_1^2\epsilon_0$, the diversity order achieved by FD-LE is one, where $\epsilon_0 = 2^{R_0} - 1$. Otherwise, the outage probability is always one.

Proof: Please refer to Appendix C. □

Remark 3: Recall that the diversity order achieved by OTFS-OMA, where the high-mobility user, U_0, solely occupies the bandwidth resources, is also one. Therefore, the use of OTFS-NOMA ensures that the additional M low-mobility users are served without compromising U_0's diversity order, which improves the spectral efficiency compared to OTFS-OMA.

B. Design and Performance of FD-DFE

Different from FD-LE, which is a linear equalizer, FD-DFE is based on the idea of feeding back previously detected symbols. Since both $\mathbf{x}_0$ and $\mathbf{x}_q$, $q \geq 1$, experience the same fading channel, we first define $\mathbf{x} = \gamma_0\mathbf{x}_0 + \sum_{q=1}^{M}\gamma_q\mathbf{x}_q$, which are the signals to be recovered by FD-DFE. Given the received signal vector shown in (12), the outputs of FD-DFE are given by

$$\hat{\mathbf{x}} = \mathbf{P}_0\mathbf{y}_0 - \mathbf{G}_0\check{\mathbf{x}}, \quad (20)$$

where $\check{\mathbf{x}}$ contains the decisions made on the symbols $\mathbf{x}$, $\mathbf{P}_0$ is the feed-forward part of the equalizer, and $\mathbf{G}_0$ is the feedback part of the equalizer. Similar to [34], [35], we use the following choices for $\mathbf{P}_0$ and $\mathbf{G}_0$: $\mathbf{P}_0 = \mathbf{L}_0(\mathbf{H}_0^H\mathbf{H}_0)^{-1}\mathbf{H}_0^H$, $\mathbf{G}_0 = \mathbf{L}_0 - \mathbf{I}_{NM}$, where $\mathbf{L}_0$ is a lower triangular matrix with its main diagonal elements being ones in order to ensure causality of the feedback signals. With the above choices for $\mathbf{P}_0$ and $\mathbf{G}_0$, U_0's signals can be detected as follows:

$$\hat{\mathbf{x}} = \mathbf{L}_0(\mathbf{H}_0^H\mathbf{H}_0)^{-1}\mathbf{H}_0^H\mathbf{y}_0 - (\mathbf{L}_0 - \mathbf{I}_{NM})\check{\mathbf{x}}. \tag{21}$$

For FD-DFE, $\mathbf{L}_0$ is obtained from the Cholesky decomposition of $\mathbf{H}_0$, i.e., $\mathbf{H}_0^H\mathbf{H}_0 = \mathbf{L}_0^H\boldsymbol{\Lambda}_0\mathbf{L}_0$, where $\mathbf{L}_0$ is the desirable lower triangular matrix, and $\boldsymbol{\Lambda}_0$ is a diagonal matrix. Therefore, the estimates of $\mathbf{x}_0$ can be rewritten as follows:

$$\hat{\mathbf{x}} = \mathbf{x} + \mathbf{L}_0(\mathbf{H}_0^H\mathbf{H}_0)^{-1}\mathbf{H}_0^H\mathbf{z}_0 \tag{22}$$

$$= \gamma_0\mathbf{x}_0 + \underbrace{\sum_{q=1}^{M}\gamma_q\mathbf{x}_q + \mathbf{L}_0(\mathbf{H}_0^H\mathbf{H}_0)^{-1}\mathbf{H}_0^H\mathbf{z}_0}_{\text{Interference and noise terms}}, \tag{23}$$

where perfect decision-making is assumed, i.e., $\check{\mathbf{x}} = \mathbf{x}$, and there is no error propagation [35], [37], [38]. We note that (23) yields an upper bound on the reception reliability of FD-DFE when error propagation cannot be completely avoided.

Following steps similar to those in the proof of Lemma 1, the covariance matrix for the interference-plus-noise term can be found as follows:

$$\mathbf{C}_{\text{cov}} = \rho\gamma_1^2\mathbf{I}_{MN} + \mathbf{L}_0(\mathbf{H}_0^H\mathbf{H}_0)^{-1}\mathbf{L}_0^H = \rho\gamma_1^2\mathbf{I}_{MN} + \boldsymbol{\Lambda}_0^{-1}, \tag{24}$$

where the last step follows from the fact that $\mathbf{L}_0$ is obtained from the Cholesky decomposition of $\mathbf{H}_0$. Therefore, the SINR for detecting $x_0[k,l]$ can be expressed as follows:

$$\text{SINR}_{0,kl} = \frac{\rho\gamma_0^2}{\rho\gamma_1^2 + \lambda_{0,kl}^{-1}}, \tag{25}$$

where $\lambda_{0,kl}$ is the $(kM{+}l{+}1)$-th element on the main diagonal of $\boldsymbol{\Lambda}_0$.

Remark 4: We note that there is a fundamental difference between the two equalization schemes. One can observe from (19) that the SINRs achieved by FD-LE for different $x_0[k,l]$ are identical. However, for FD-DFE, different symbols experience different effective fading gains, $\lambda_{0,kl}$. Therefore, FD-DFE can realize unequal error protection for data streams with different priorities. This comes at the price of a higher computational complexity.

We further note that the use of FD-DFE also ensures that multi-path diversity can be harvested, as shown in the following. The outage performance analysis for FD-DFE requires knowledge of the distribution of the effective channel gains, $\lambda_{0,kl}$. Because of the implicit relationship between $\boldsymbol{\Lambda}_0$ and $\mathbf{H}_0$, a general expression for the outage probability achieved by FD-DFE is difficult to obtain. However, analytical results can be developed for special cases to show that the use of FD-DFE can realize the maximal multi-path diversity.

In particular, the SINR for $x_0[N-1,M-1]$ is a function of $\lambda_{0,(N-1)(M-1)}$ which is the last element on the main diagonal of $\boldsymbol{\Lambda}_0$. Recall that $\boldsymbol{\Lambda}_0$ is obtained via Cholesky decomposition, i.e., $\mathbf{H}_0^H\mathbf{H}_0 = \mathbf{L}_0^H\boldsymbol{\Lambda}_0\mathbf{L}_0$. Because $\mathbf{L}_0$ is a lower triangular matrix, $\lambda_{0,(N-1)(M-1)}$ is equal to the element of $\mathbf{H}_0^H\mathbf{H}_0$ located in the NM-th column and the NM-th row, which means

$$\lambda_{0,(N-1)(M-1)} = \sum_{p=0}^{P_0}|h_{0,p}|^2. \tag{26}$$

Since the channel gains are i.i.d. and follow $h_{0,p} \sim CN(0,\frac{1}{P_0+1})$, the probability density function (pdf) of $\sqrt{P_0+1}\lambda_{0,(N-1)(M-1)}$ is given by

$$f(x) = \frac{1}{P_0!}e^{-x}x^{P_0}. \tag{27}$$

By using the above pdf, the outage probability and the diversity order can be obtained by some algebraic manipulations, as shown in the following corollary.

Corollary 1: Assume $\gamma_0^2 > \gamma_1^2\epsilon_0$. The use of FD-DFE realizes the following outage probability for detection of $x_0[N-1,M-1]$:

$$\mathrm{P}_{N-1,M-1}^0 = \frac{1}{P_0!}g\left(P_0+1, \frac{\epsilon_0(P_0+1)}{\rho(\gamma_0^2-\gamma_1^2\epsilon_0)}\right), \tag{28}$$

where $g(\cdot)$ denotes the incomplete Gamma function. The full multi-path diversity order, P_0+1, is achievable for $x_0[N-1, M-1]$.

Remark 5: The results in Corollary 1 can be extended to OTFS-OMA with FD-DFE straightforwardly. We also note that diversity gains larger than one are not achievable with FD-LE as shown in Lemma 2, which is one of the disadvantages of FD-LE compared to FD-DFE.

Remark 6: We note that not all NM data streams can benefit from the full diversity gain. The simulation results provided in Section VII (Fig. 2) show that the diversity orders achievable for $x_0[k,l]$, $k < N-1$ and $l < M-1$, are smaller than that for $x_0[N-1,M-1]$, and the diversity order for $x_0[0,0]$ is one, i.e., the same value as for FD-LE. We further note that the diversity result in Corollary 1 is obtained by assuming that there is no error propagation, i.e., it is assumed that when detecting the i-th element of $\mathbf{x}$ in (21), the first $(i-1)$ elements of $\mathbf{x}$ have already been correctly detected. Because of this assumption, the diversity gain developed in Corollary 1 is an upper bound on the diversity gain achieved by FD-DFE. If the assumption does not hold, the diversity orders for $x_0[k,l]$ will be capped by the worst case, i.e., the diversity gain for $x_0[0,0]$ which is one.

Remark 7: FD-DFE entails a higher implementation complexity than FD-LE, as explained in the following. The complexity of FD-LE is mainly caused by computing the inversion of $\mathbf{H}_0^H\mathbf{H}_0$. However, for FD-DFE, $\mathbf{L}_0$ needs to be computed, in addition to $(\mathbf{H}_0^H\mathbf{H}_0)^{-1}$, as shown in (21). Recall that $\mathbf{L}_0$ is obtained from the Cholesky decomposition of the $NM \times NM$ matrix $\mathbf{H}_0$, which entails a computational complexity of $\mathcal{O}(N^3M^3)$. Therefore, the computational complexity of FD-DFE is higher than that of FD-LE, but FD-DFE offers a performance gain in terms of reception reliability compared to FD-LE, as shown in Section VII.

V. Downlink OTFS-NOMA - Detecting the NOMA Users' Signals

Successive interference cancellation (SIC) will be carried out by the NOMA users, where each NOMA user first decodes the high mobility user's signal in the delay-Doppler plane and then decodes its own signal in the time-frequency plane. The two stages of SIC are discussed in the following two subsections, respectively.

A. Stage I of SIC

Following steps similar to the ones in the previous section, each NOMA user also observes the mixture of the $(M+1)$ users' signals in the delay-Doppler plane as follows:

$$\mathbf{y}_i = \gamma_0 \mathbf{H}_i \mathbf{x}_0 + \underbrace{\sum_{q=1}^{M} \gamma_q \mathbf{H}_i \mathbf{x}_q + \mathbf{z}_i}_{\text{Interference and noise terms}}, \tag{29}$$

where $\mathbf{H}_i$ and $\mathbf{z}_i$ are defined similar to $\mathbf{H}_0$ and $\mathbf{z}_0$, respectively.

We assume that the low-mobility NOMA users do not experience Doppler shift, and therefore, their channels can be simplified as follows:

$$h_i(\tau) = \sum_{p=0}^{P_i} h_{i,p} \delta(\tau - \tau_{i,p}), \tag{30}$$

for $1 \leq i \leq K$, which means that each NOMA user's channel matrix, $\mathbf{H}_i$, $1 \leq i \leq N$, is a block-diagonal matrix, i.e., $\mathbf{A}_{i,0}$ is a non-zero circulant matrix and $\mathbf{A}_{i,n} = \mathbf{0}_{M\times M}$, for $1 \leq n \leq N-1$. Therefore, each NOMA user can divide its observation vector into N equal-length sub-vectors, i.e., $\mathbf{y}_i = \left[\mathbf{y}_{i,0}^T \cdots \mathbf{y}_{i,N-1}^T\right]^T$, which yields the following simplified system model:

$$\mathbf{y}_{i,n} = \gamma_0 \mathbf{A}_{i,0} \mathbf{x}_{0,n} + \sum_{q=1}^{M} \gamma_q \mathbf{A}_{i,0} \mathbf{x}_{q,n} + \mathbf{z}_{i,n}, \tag{31}$$

where, similar to $\mathbf{y}_{i,n}$, $\mathbf{x}_{i,n}$ and $\mathbf{z}_{i,n}$ are obtained from $\mathbf{x}_i$ and $\mathbf{z}_i$, respectively. Therefore, unlike the high-mobility user, the NOMA users can perform their signal detection based on reduced-size observation vectors, which reduces the computational complexity.

Since $\mathbf{A}_{i,0}$ is a circulant matrix, the two equalization approaches used in the previous section are still applicable. First, we consider the use of FD-LE. Following the same steps as in the proof for Proposition 1, in the first step of FD-LE, the DFT matrix is applied to the reduced-size observation vector, which yields the following:

$$\tilde{\mathbf{y}}_{i,n} = \tilde{\mathbf{D}}_i \mathbf{F}_M^H \left(\gamma_0 \mathbf{x}_{0,n} + \sum_{q=1}^{M} \gamma_q \mathbf{x}_{q,n} \right) + \tilde{\mathbf{z}}_{i,n}, \tag{32}$$

where $\tilde{\mathbf{y}}_{i,n} = \mathbf{F}_M^H \mathbf{y}_{i,n}$ and $\tilde{\mathbf{z}}_{i,n} = \mathbf{F}_M^H \mathbf{z}_{i,n}$. Compared to $\mathbf{D}_i$ in Proposition 1 which is an $NM \times NM$ matrix, $\tilde{\mathbf{D}}_i$ is an $M \times M$ diagonal matrix, and its $(l+1)$-th main diagonal element is given by $\tilde{D}_i^l = \sum_{m=0}^{M-1} a_{i,0}^{m,1} e^{j2\pi \frac{lm}{M}}$, for $0 \leq l \leq M-1$, where $a_{i,0}^{m,1}$ is the element located in the $(m+1)$-th row and the first column of $\mathbf{A}_{i,0}$. Unlike conventional OFDM, which uses $\mathbf{F}_M$ at the receiver, $\mathbf{F}_M^H$ is used here. Because $\mathbf{F}_M^H \mathbf{A}_{i,0} \mathbf{F}_M = \left[\mathbf{F}_M \mathbf{A}_{i,0}^* \mathbf{F}_M^H\right]^*$, the sign of the exponent of the exponential component of $\tilde{D}_i^l$ is different from that in the conventional case.

In the second step of FD-LE, $\mathbf{F}_M \tilde{\mathbf{D}}_i^{-1}$ is applied to $\tilde{\mathbf{y}}_{i,n}$. Following steps similar to the ones in the proof for Lemma 1, the SINR for detecting $x_0[k,l]$ can be obtained as follows:

$$\text{SINR}_{0,kl}^{i,\text{LE}} = \frac{\rho\gamma_0^2}{\rho\gamma_1^2 + \frac{1}{M}\sum_{\tilde{l}=0}^{M-1} |\tilde{D}_i^{\tilde{l}}|^{-2}}. \tag{33}$$

We note that $\text{SINR}_{0,k_1 l}^{i,\text{LE}} = \text{SINR}_{0,k_2 l}^{i,\text{LE}}$, for $k_1 \neq k_2$, due to the time invariant nature of the channels.

If FD-DFE is used, the corresponding SINR for detecting $x_0[k,l]$ is given by

$$\text{SINR}_{0,kl}^{i,\text{DFE}} = \frac{\rho\gamma_0^2}{\rho\gamma_1^2 + \tilde{\lambda}_{0,l}^{-1}}, \tag{34}$$

where $\tilde{\lambda}_{0,l}$ is obtained from the Cholesky decomposition of $\mathbf{A}_{i,0}$. The details for the derivation of (34) are omitted here due to space limitations.

B. Stage II of SIC

Assume that U_0's NM signals can be decoded and removed successfully, which means that, in the time-frequency plane, the NOMA users observe the following:

$$\begin{aligned} Y_i[n,m] &= \sum_{q=1}^{M} \gamma_q H_i[n,m] X_q[n,m] + W_i[n,m] \\ &= \gamma_1 H_i[n,m] x_{m+1}(n) + W_i[n,m], \end{aligned} \tag{35}$$

where the last step follows from the mapping scheme used in (6) and it is assumed that all NOMA users employ the same power allocation coefficient. We note that U_i is only interested in $Y_i[n,i-1]$, $0 \leq n \leq N-1$. Therefore, U_i's n-th information bearing signal, $x_i(n)$, can be detected by applying a one-tap equalizer as follows:

$$\hat{x}_i(n) = \frac{Y_i[n,i-1]}{\gamma_1 H_i[n,i-1]}, \tag{36}$$

which means that the SNR for detecting $x_i(n)$ is given by

$$\text{SNR}_{i,n} = \rho\gamma_1^2 |\tilde{D}_i^{i-1}|^2, \tag{37}$$

since $W_i[n,i-1]$ is white Gaussian noise and $H_i[n,i-1] = \tilde{D}_i^{i-1}$. We note that $\text{SNR}_{i,n_1} = \text{SNR}_{i,n_2}$, for $n_1 \neq n_2$, which is due to the time-invariant nature of the channel.

Without loss of generality, assume that the same target data rate R_i is used for $x_i(n)$, $0 \leq n \leq N-1$. Therefore, the outage probability for $x_i(n)$ is given by

$$\begin{aligned} &\text{P}_{i,n}^{\text{LE}} \\ &= 1 - \text{P}\left(\text{SNR}_{i,n} > \epsilon_i, \text{SINR}_{0,kl}^{i,\text{LE}} > \epsilon_0, \forall l\right) \\ &= 1 - \text{P}\left(\rho\gamma_1^2 |\tilde{D}_i^{i-1}|^2 > \epsilon_i, \frac{\rho\gamma_0^2}{\rho\gamma_1^2 + \frac{1}{M}\sum_{l=0}^{M-1} |\tilde{D}_i^l|^{-2}} > \epsilon_0\right), \end{aligned} \tag{38}$$

if FD-LE is used in the first stage of SIC. If FD-DFE is used in the first stage of SIC, the outage probability for $x_i(n)$ is given by

$$\begin{aligned} \text{P}_{i,n}^{\text{DFE}} &= 1 - \text{P}\left(\text{SNR}_{i,n} > \epsilon_i, \text{SINR}_{0,kl}^{i,\text{DFE}} > \epsilon_0, \forall l\right) \\ &= 1 - \text{P}\left(\rho\gamma_1^2 |\tilde{D}_i^{i-1}|^2 > \epsilon_i, \frac{\rho\gamma_0^2}{\rho\gamma_1^2 + \tilde{\lambda}_{0,l}^{-1}} > \epsilon_0, \forall l\right), \end{aligned} \tag{39}$$

where $\epsilon_i = 2^{R_i} - 1$. Again because of the correlation between the random variables $|\tilde{D}_i^l|^{-2}$ and $\tilde{\lambda}_{0,l}$, the exact expressions for the outage probabilities are difficult to obtain. Alternatively, the achievable diversity order is analyzed in the following subsections.

1) Random User Scheduling: If the M users are randomly selected from the K available users, which means that each $|\tilde{D}_i^l|^2$ is complex Gaussian distributed. For the FD-LE case, the outage probability, $\mathrm{P}_{i,n}^{\mathrm{LE}}$, can be upper bounded as follows:

$$\mathrm{P}_{i,n}^{\mathrm{LE}} \leq 1 - \mathrm{P}\left(\rho\gamma_1^2|\tilde{D}_i^{\min}|^2 > \epsilon_i, \frac{\rho\gamma_0^2}{\rho\gamma_1^2 + |\tilde{D}_i^{\min}|^{-2}} > \epsilon_0\right), \tag{40}$$

where $|\tilde{D}_i^{\min}|^2 = \min\{|\tilde{D}_i^m|^2, 0 \leq m \leq M-1\}$. The upper bound on the outage probability in (40) can be rewritten as follows:

$$\mathrm{P}_{i,n}^{\mathrm{LE}} \leq 1 - \mathrm{P}\left(|\tilde{D}_i^{\min}|^2 > \bar{\epsilon}\right), \tag{41}$$

where $\bar{\epsilon} = \max\left\{\frac{\epsilon_0}{\rho(\gamma_0^2-\gamma_1^2\epsilon_0)}, \frac{\epsilon_i}{\rho\gamma_1^2}\right\}$. As a result, an upper bound on the outage probability can be obtained as follows:

$$\mathrm{P}_{i,n}^{\mathrm{LE}} \leq \mathrm{P}\left(|\tilde{D}_i^{\min}|^2 < \bar{\epsilon}\right) \leq M\mathrm{P}\left(|\tilde{D}_i^0|^2 < \bar{\epsilon}\right) \doteq \frac{1}{\rho}, \tag{42}$$

where $\mathrm{P}^o \doteq \rho^{-d}$ denotes exponential equality, i.e., $d = -\lim_{\rho\to\infty} \frac{\log \mathrm{P}^o}{\log\rho}$ [36]. Therefore, the following corollary can be obtained.

Corollary 2: For random user scheduling and FD-LE, a diversity order of 1 is achievable at the NOMA users.

Our simulation results in Section VII show that a diversity order of 1 is also achievable for FD-DFE, although we do not have a formal proof for this conclusion, yet.

2) Realizing Multi-User Diversity: The diversity order of OTFS-NOMA can be improved by carrying out opportunistic user scheduling, which yields multi-user diversity gains. For illustration purpose, we propose a greedy user scheduling policy, where a single NOMA user is scheduled to transmit in all resource blocks of the time-frequency plane. From the analysis of the random scheduling case we deduce that $|\tilde{D}_i^{\min}|^2$ is critical to the outage performance. Therefore, the scheduled NOMA user, denoted by U_{i^*}, is selected based on the following criterion:

$$i^* = \arg\max_{i\in\{1,\cdots,K\}} \left\{|\tilde{D}_i^{\min}|^2\right\}. \tag{43}$$

By using the assumption that the users' channel gains are independent and following steps similar to the ones in the proof for Lemma 2, the following corollary can be obtained in a straightforward manner.

Corollary 3: For FD-LE, the user scheduling strategy shown in (43) *realizes the maximal multi-user diversity gain, K.*

Remark 8: The reason why a multi-user diversity gain of K can be realized by the proposed scheduling strategy is explained in the following. Recall that the SINR for FD-LE to detect $x_0[k,l]$ is $\mathrm{SINR}_{0,kl}^{i,\mathrm{LE}} = \frac{\rho\gamma_0^2}{\rho\gamma_1^2 + \frac{1}{M}\sum_{\tilde{l}=0}^{M-1}|\tilde{D}_i^{\tilde{l}}|^{-2}}$. If this SINR is too small, the first stage of SIC will fail and an outage event will occur. To improve the SINR, it is important to ensure that for a scheduled user, its weakest channel gain, $|\tilde{D}_i^{\min}|^2 = \min\{|\tilde{D}_i^m|^2, 0 \leq m \leq M-1\}$, is not too small. The used scheduling strategy shown in (43) is essentially a max-min strategy and ensures that the user with the strongest $|\tilde{D}_i^{\min}|^2$ is selected from the K candidates, which effectively exploits multi-user diversity.

We note that the user scheduling strategy shown in (43) is also useful for improving the performance of FD-DFE, as shown in Section VII.

VI. Uplink OTFS-NOMA Transmission

The design of uplink OTFS-NOMA is similar to that of downlink OTFS-NOMA, and due to space limitations, we mainly focus on the difference between the two cases in this section. Again, we assume that U_0 is grouped with M NOMA users, selected from the K available users. U_0's NM signals are placed in the delay-Doppler plane, and are denoted by $x_0[k,l]$, where $0 \leq k \leq N-1$ and $0 \leq l \leq M-1$. The corresponding time-frequency signals, $X_0[n,m]$, are obtained by applying ISFFT to $x_0[k,l]$. On the other hand, the NOMA users' signals, $x_i(n)$, are mapped to time-frequency signals, $X_i[n,m]$, according to (6).

Following steps similar to the ones for the downlink case, the base station's observations in the time-frequency plane are given by

$$\begin{aligned} Y[n,m] &= \sum_{q=0}^{M} H_q[n,m]X_q[n,m] + W[n,m] \\ &= \frac{H_0(n,m)}{NM}\sum_{k=0}^{N-1}\sum_{l=0}^{M-1} x_0[k,l]e^{j2\pi\left(\frac{kn}{N}-\frac{ml}{M}\right)} \\ &\quad + \sum_{q=1}^{M} H_q[n,m]X_q[n,m] + W[n,m], \end{aligned} \tag{44}$$

where $W[n,m]$ is the Gaussian noise at the base station in the time-frequency plane. We assume that all users employ the same transmit pulse as well as the same transmit power. The base station applies SIC to first detect the NOMA users' signals in the time-frequency plane, and then tries to detect the high-mobility user's signals in the delay-Doppler plane, as shown in the following two subsections.

A. Stage I of SIC

The base station will first try to detect the NOMA users' signals in the time-frequency plane by treating the signals from U_0 as noise, which is the first stage of SIC.

By using (6), $x_i(n)$ can be estimated as follows:

$$\begin{aligned} \hat{x}_i(n) &= \frac{Y[n,i-1]}{H_i[n,i-1]} \\ &= x_i[n] + \frac{H_0[n,i-1]X_0[n,i-1] + W[n,i-1]}{H_i[n,i-1]}. \end{aligned} \tag{45}$$

Define an $NM\times 1$ vector, $\bar{\mathbf{x}}_0$, whose $(nM+m+1)$-th element is $X_0[n,m]$. Recall that $X_0[n,m]$ is obtained from the ISFFT of $x_0[k,l]$, i.e.,

$$\bar{\mathbf{x}}_0 = (\mathbf{F}_N^H \otimes \mathbf{F}_M)\mathbf{x}_0, \tag{46}$$

which means $X_0[n,m]$ follows the same distribution as $x_0[k,l]$. By applying steps similar to those in the proof for Lemma 1, the SINR for detecting $x_i(n)$ is given by

$$\mathrm{SINR}_{i,n} = \frac{\rho|H_i[n,i-1]|^2}{\rho|H_0[n,i-1]|^2+1}. \tag{47}$$

Unlike downlink OTFS-NOMA, there are two possible strategies for uplink OTFS-NOMA to combat multiple access interference, as shown in the following two subsections.

1) Adaptive-Rate Transmission: One strategy to combat multiple access interference is to impose the following constraint on $x_i(n)$:

$$R_{i,n} \leq \log\left(1+\frac{\rho|H_i[n,i-1]|^2}{\rho|H_0[n,i-1]|^2+1}\right), \tag{48}$$

which means that the first stage of SIC is guaranteed to be successful. Therefore, the M low-mobility users are served without affecting U_0's outage probability, i.e., the use of NOMA is transparent to U_0.

Because U_i's data rate is adaptive, outage events when decoding $x_i(n)$ do not happen, which means that an appropriate criterion for the performance evaluation is the ergodic rate. Recall that $H_i[n,i-1] = \tilde{D}_i^{i-1}$ and $H_0[n,i-1] = D_0^{n,i-1}$. Therefore, U_i's ergodic rate is given by

$$\mathcal{E}\{R_{i,n}\} \leq \mathcal{E}\left\{\log\left(1+\frac{\rho|\tilde{D}_i^{i-1}|^2}{\rho|D_0^{n,i-1}|^2+1}\right)\right\}. \tag{49}$$

We note that the ergodic rate of uplink OTFS-NOMA can be further improved by modifying the user scheduling strategy proposed in (43), as shown in the following. Particularly, denote the NOMA user which is scheduled to transmit in the m-th frequency subchannel by $\mathrm{U}_{i_m^*}$, and this user is selected by using the following criterion:

$$i_m^* = \arg \max_{i\in\{1,\cdots,K\}} \left\{|\tilde{D}_i^m|^2\right\}. \tag{50}$$

We note that a single user might be scheduled on multiple frequency channels, which reduces user fairness.

Because the integration of the logarithm function appearing in (49) leads to non-insightful special functions, we will use simulations to evaluate the ergodic rate of OTFS-NOMA in Section VII.

2) Fixed-Rate Transmission: If the NOMA users do not have the capabilities to adapt their transmission rates, they have to use fixed data rates R_i for transmission, which means that outage events can happen and the achieved outage performance is analyzed in the following. For illustration purposes, we focus on the case when the user scheduling strategy shown in (50) is used.

The outage probability for detecting $x_{i_m^*}(n)$ is given by

$$\mathrm{P}_{i_m^*,n} = \mathrm{P}\left(\log\left(1+\frac{\rho|\tilde{D}_{i_m^*}^{i_m^*-1}|^2}{\rho|D_0^{n,i_m^*-1}|^2+1}\right) < R_{i_m^*}\right). \tag{51}$$

Following steps similar to the ones in the proof for Lemma 2, we can show that $|\tilde{D}_{i_m^*}^{i_m^*-1}|^2$ and $|D_0^{n,i_m^*-1}|^2$ are independent, and the use of the user scheduling scheme in (50) simplifies the outage probability as follows:

$$\begin{aligned}\mathrm{P}_{i_m^*,n} &= \mathrm{P}\left(\log\left(1+\frac{\rho|\tilde{D}_{i_m^*}^{i_m^*-1}|^2}{\rho|D_0^{n,i_m^*-1}|^2+1}\right) < R_{i_m^*}\right)\\ &= \int_0^\infty \left(1-e^{-\frac{\epsilon_{i_m^*}(1+\rho y)}{\rho}}\right)^K e^{-y}dy, \end{aligned}\tag{52}$$

where we use the fact that the cumulative distribution function of $|\tilde{D}_{i_m^*}^{i_m^*-1}|^2$ is $(1-e^{-x})^K$ because of the adopted user scheduling strategy.

The outage probability can be further simplified as follows:

$$\begin{aligned}\mathrm{P}_{i_m^*,n} &= \sum_{k=0}^K \binom{K}{k}(-1)^k \int_0^\infty e^{-\frac{k\epsilon_{i_m^*}(1+\rho y)}{\rho}-y}dy\\ &= \sum_{k=0}^K \binom{K}{k}(-1)^k e^{-\frac{k\epsilon_{i_m^*}}{\rho}} \frac{1}{k\epsilon_{i_m^*}+1}. \end{aligned}\tag{53}$$

At high SNR, the outage probability can be approximated as follows:

$$\mathrm{P}_{i_m^*,n} \approx \sum_{k=0}^K \binom{K}{k}(-1)^k \frac{1}{k\epsilon_{i_m^*}+1}, \tag{54}$$

which is no longer a function of ρ, i.e., the outage probability has an error floor at high SNR. This is due to the fact that $\mathrm{U}_{i_m^*}$ is subject to strong interference from U_0.

However, we can show that the error floor experienced by $\mathrm{U}_{i_m^*}$ can be reduced by increasing K, i.e., inviting more opportunistic users for NOMA transmission. In particular, assuming $K\epsilon_{i_m^*} \to 0$, the outage probability can be approximated as follows:

$$\begin{aligned}\mathrm{P}_{i_m^*,n} &\approx \sum_{k=0}^K \binom{K}{k}(-1)^k \left(1+k\epsilon_{i_m^*}\right)^{-1}\\ &\approx \sum_{k=0}^K \binom{K}{k}(-1)^k \sum_{l=0}^\infty (-1)^l k^l \epsilon_{i_m^*}^l, \end{aligned}\tag{55}$$

where we use the fact that $(1+x)^{-1} = \sum_{l=0}^\infty (-1)^l x^l$, $|x|<1$. Therefore, the error floor at high SNR can be approximated as follows:

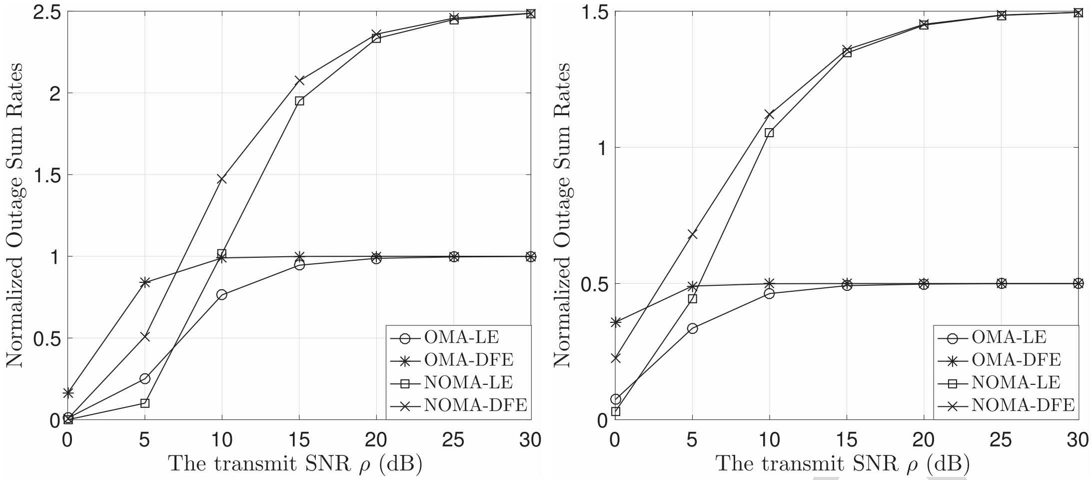
$$\begin{aligned}\mathrm{P}_{i_m^*,n} &\approx \sum_{l=0}^\infty (-1)^l \epsilon_{i_m^*}^l \sum_{k=0}^K \binom{K}{k}(-1)^k k^l\\ &\approx (-1)^K \epsilon_{i_m^*}^K (-1)^K K! = K!\epsilon_{i_m^*}^K, \end{aligned}\tag{56}$$

where we use the identities $\sum_{k=0}^K \binom{K}{k}(-1)^k k^l = 0$, for $l<K$ and $\sum_{k=0}^K \binom{K}{k}(-1)^k k^K = (-1)^K K!$.

The conclusion that increasing K reduces the error floor can be confirmed by defining $f(k) = k!\epsilon_{i_m^*}^k$ and using the following fact:

$$f(k) - f(k+1) = k!\epsilon_{i_m^*}^k \left(1-(k+1)\epsilon_{i_m^*}\right) > 0, \tag{57}$$

where it is assumed that $k\epsilon_{i_m^*} \to 0$.



(a) $R_0 = 1$ BPCU and $R_i = 1.5$ BPCU (b) $R_0 = 0.5$ BPCU and $R_i = 1$ BPCU

Fig. 1. Impact of OTFS-NOMA on the downlink sum rates. $M = N = K = 16$. $P_0 = P_i = 3$. BPCU denotes bit per channel use. $\gamma_0^2 = \frac{3}{4}$ and $\gamma_i^2 = \frac{1}{4}$ for $i > 0$. Random user scheduling is used.

B. Stage II of SIC

If adaptive transmission is used, the NOMA users' signals can be detected successfully during the first stage of SIC. Therefore, they can be removed from the observations at the base station, i.e., $\bar{Y}[n,m] = Y[n,m] - \sum_{q=1}^{N} H_q(n,m)X_q[n,m]$, and SFFT is applied to obtain the delay-Doppler observations as follows:

$$\begin{aligned} y_0[k,l] &= \frac{1}{NM}\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}\bar{Y}[n,m]e^{-j2\pi\left(\frac{nk}{N}-\frac{ml}{M}\right)} \\ &= \sum_{p=0}^{P_0} h_{0,p}x_0[(k-k_{\mu_{0,p}})_N,(l-l_{\tau_{0,p}})_M]+z[k,l], \quad (58) \end{aligned}$$

where $z[k,l]$ denote additive noise. U_0's signals can be detected by applying either of the two considered equalization approaches, and the same performance as for OTFS-OMA can be realized. The analytical development is similar to the downlink case, and hence is omitted due to space limitations.

However, if fixed-rate transmission is used, the uplink outage events for decoding $x_0[k,l]$ are different from the downlink ones, as shown in the following. Particularly, the use of FD-LE yields the following SINR expression for decoding $x_0[k,l]$:

$$\mathrm{SINR}_{0,kl}^{\mathrm{LE}} = \frac{\rho}{\frac{1}{NM}\sum_{k=0}^{N-1}\sum_{l=0}^{M-1}|D_0^{k,l}|^{-2}}. \quad (59)$$

If FD-DFE is used, the SNR for detection of $x_0[k,l]$ is given by

$$\mathrm{SINR}_{0,kl}^{\mathrm{DFE}} = \rho\lambda_{0,kl}. \quad (60)$$

Therefore, the outage probability for detecting $x_0[k,l]$ is given by

$$\begin{aligned} \mathrm{P}_{kl} &= 1 - \mathrm{P}\left(\mathrm{SINR}_{0,kl}^{\mathrm{DFE/LE}} > \epsilon_0, \mathrm{SNR}_{i,n} > \epsilon_i \forall i,n\right) \\ &\geq 1 - \mathrm{P}\left(\mathrm{SNR}_{i,n} > \epsilon_i \forall i,n\right) \geq \mathrm{P}\left(\mathrm{SNR}_{1,0} < \epsilon_i\right). \end{aligned}$$

Since $\mathrm{P}\left(\mathrm{SNR}_{1,0} < \epsilon_i\right)$ has an error floor as shown in the previous subsection, the uplink outage probability for detection of U_0's signals does not go to zero even if $\rho \to \infty$, which is different from the downlink case. Therefore, if fixed-rate transmission is used, adding the M low-mobility users into the bandwidth, which would be solely occupied by U_0 in OTFS-OMA, improves connectivity but degrades U_0's performance.

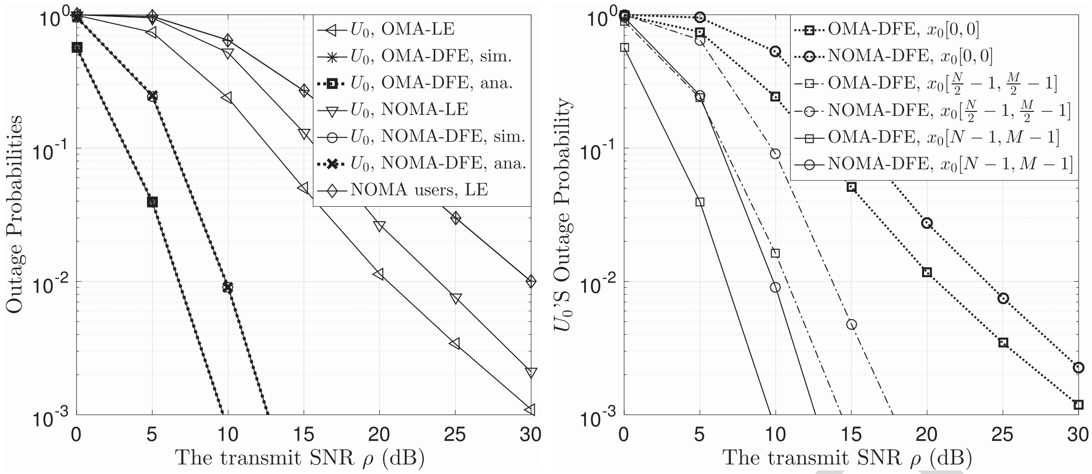
TABLE I
DELAY-DOPPLER PROFILE FOR U_0'S CHANNEL

Propagation path index (p)	0	1	2	3
Delay ($\tau_{0,p}$) μs	8.33	25	41.67	58.33
Delay tap index ($l_{\tau_{0,p}}$)	2	6	10	14
Doppler ($\nu_{0,p}$) Hz	0	0	468.8	468.8
Doppler tap index ($k_{\nu_{0,p}}$)	0	0	1	1

VII. NUMERICAL STUDIES

In this section, the performance of OTFS-NOMA is evaluated via computer simulations. Similar to [26]–[28], we first define the delay-Doppler profile for U_0's channel as shown in Table I, where $P_0 = 3$ and the subchannel spacing is $\Delta f = 7.5$ kHz. Therefore, the maximal speed corresponding to the largest Doppler shift $\nu_{0,3} = 468.8$ Hz is 126.6 km/h if the carrier frequency is $f_c = 4$ GHz. On the other hand, the NOMA users' channels are assumed to be time invariant with $P_i = 3$ propagation paths, i.e., $\tau_{i,p} = 0$ for $p \geq 4$, $i \geq 1$. For all the users' channels, we assume that $\sum_{p=0}^{P_i}\mathcal{E}\{|h_{i,p}|^2\} = 1$ and $|h_{i,p}|^2 \sim CN\left(0, \frac{1}{P_i+1}\right)$. For the fixed rate transmission scheme, a simple choice for power allocation ($\gamma_0^2 = \frac{3}{4}$ and $\gamma_i^2 = \frac{1}{4}$ for $i > 0$) is considered. The performance of OTFS-NOMA could be further improved by optimizing γ_i according to the users' channel conditions and QoS requirements.

In Fig. 1, downlink OTFS-NOMA transmission is evaluated by using the normalized outage sum rate as the performance criterion which is defined as $\frac{1}{NM}\sum_{k=0}^{N-1}\sum_{l=0}^{M-1}(1-\mathrm{P}_{0,kl})R_0$



(a) Outage probabilities of U_0 and the NOMA users (b) Performance of FD-DFE

Fig. 2. The outage performance of downlink OTFS-OMA and OTFS-NOMA. $M = N = K = 16$. $P_0 = P_i = 3$. $\gamma_0^2 = \frac{3}{4}$ and $\gamma_i^2 = \frac{1}{4}$ for $i > 0$. $R_0 = 0.5$ BPCU and $R_i = 1$ BPCU. In Fig. 2(a), for FD-DFE, the performance of $x_0[N-1, M-1]$ is shown. Random user scheduling is used.

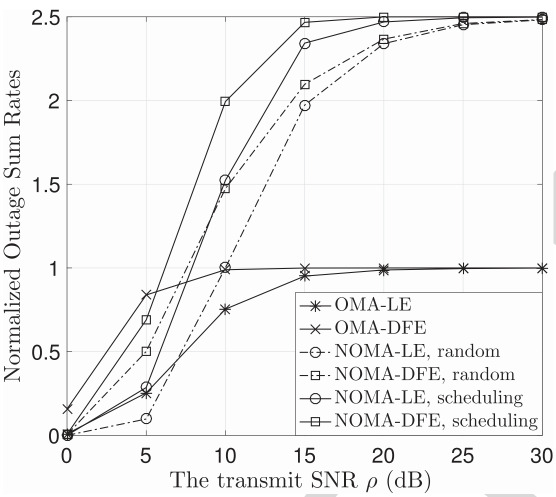


Fig. 3. Impact of user scheduling on the downlink outage sum rates. $P_0 = P_i = 3$. $R_0 = 1$ BPCU and $R_i = 1.5$ BPCU. $M = N = K = 16$, $\gamma_0^2 = \frac{3}{4}$ and $\gamma_i^2 = \frac{1}{4}$ for $i > 0$.

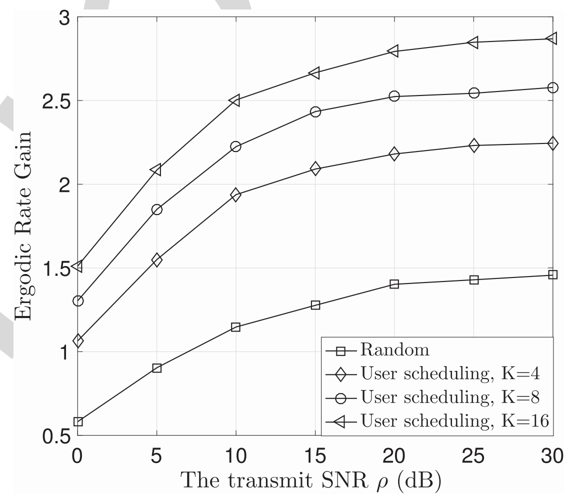
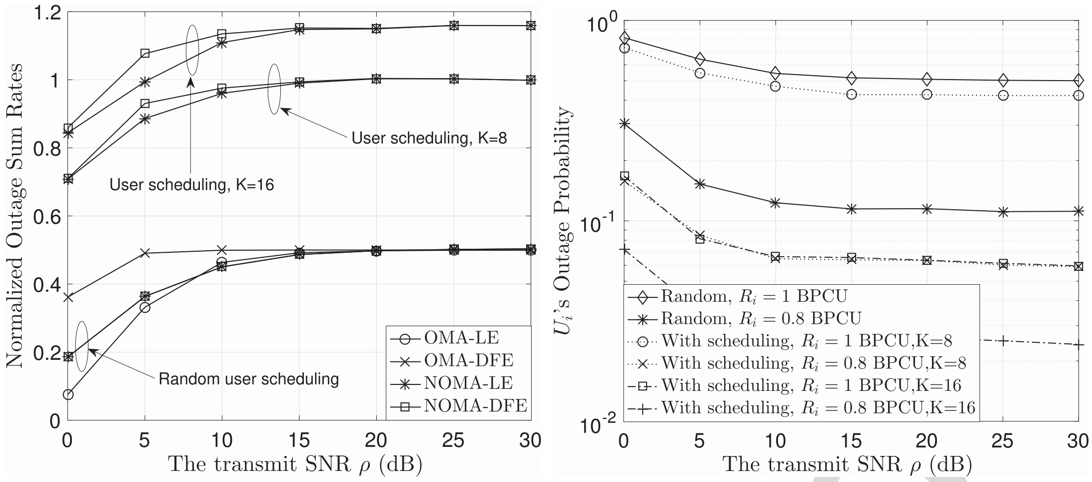


Fig. 4. The ergodic rate gain of OTFS-NOMA over OTFS-OMA. The NOMA users adapt their data rates according to (48). $P_0 = P_i = 3$. $M = N = 16$.

and $\frac{1}{NM}\sum_{k=0}^{N-1}\sum_{l=0}^{M-1}(1-\mathrm{P}_{0,kl})R_0 + \frac{1}{NM}\sum_{i=1}^{M}\sum_{n=0}^{N-1}(1-\mathrm{P}_{i,n})R_i$ for OTFS-OMA and OTFS-NOMA, respectively. Fig. 1 shows that the use of OTFS-NOMA can significantly improve the sum rate at high SNR for both considered choices of R_0 and R_i. The reason for this performance gain is the fact that the maximal sum rate achieved by OTFS-OMA is capped by R_0, whereas OTFS-NOMA can provide sum rates up to $R_0 + R_i$. Comparing Fig. 1(a) to Fig. 1(b), one can observe that the performance loss of OTFS-NOMA at low SNR can be mitigated by reducing the target data rates, since reducing the target rates improves the probability of successful SIC. Furthermore, both figures show that FD-DFE outperforms FD-LE in the entire considered range of SNRs; however, we note that the performance gain of FD-DFE over FD-LE is achieved at the expense of increased computational complexity.

In Fig. 2, the outage probabilities achieved by downlink OTFS-OMA and OTFS-NOMA are shown. As can be seen from Fig. 2(a), the diversity order achieved with FD-LE for detection of $x_0[k,l]$ is one, as expected from Lemma 2. As discussed in Section IV-B, one advantage of FD-DFE over FD-LE is that FD-DFE facilitates multi-path fading diversity gains, whereas FD-LE is limited to a diversity gain of one. This conclusion is confirmed by Fig. 2(a), where the analytical results developed in Corollary 1 are also verified. Fig. 2(b) shows the outage probabilities achieved by FD-DFE for different $x_0[k,l]$. As shown in the figure, the lowest outage probability is obtained for $x_0[N-1, M-1]$, whereas the outage probability of $x_0[0,0]$ is the largest, which is due to the fact that, in FD-DFE, different signals $x_0[k,l]$ are affected by different effective channel gains, $\lambda_{0,kl}$. Another important observation from the figures is that the FD-LE outage probability is the same as the FD-DFE outage probability for detection of $x_0[0,0]$, which fits the intuition that for FD-DFE the reliability of the first decision ($x_0[0,0]$) is the same as that of FD-LE. For the same reason, FD-LE and FD-DFE yield similar performance for detection of the NOMA users'



(a) Outage Sum Rate ($R_i = 1$ BPCU) (b) Outage Probability

Fig. 5. The performance of uplink OTFS-NOMA. Fixed-rate transmission is used by the NOMA users. $M = N = 16$. $P_0 = P_i = 3$. $R_0 = 0.5$ BPCU. $\gamma_0^2 = \frac{3}{4}$ and $\gamma_i^2 = \frac{1}{4}$ for $i > 0$.

signals, since the FD-DFE outage performance is dominated by the reliability for detection of $x_0[0, 0]$, and hence is the same as that of FD-LE.

In addition to multi-path diversity, another degree of freedom available in the considered OTFS-NOMA downlink scenario is multi-user diversity, which can be harvested by applying user scheduling as discussed in Section V-B. Fig. 3 demonstrates the benefits of exploiting multi-user diversity. With random user scheduling, at low SNR, the performance of OTFS-NOMA is worse than that of OTFS-OMA, which is also consistent with Fig. 1. By increasing the number of users participating in OTFS-NOMA, the performance of OTFS-NOMA can be improved, particularly at low and moderate SNR. For example, for FD-LE, the performance of OTFS-NOMA approaches that of OTFS-OMA at low SNR by exploiting multi-user diversity, and for FD-DFE, an extra gain of 0.5 BPCU can be achieved at moderate SNR.

In Figs. 4 and 5, the performance of uplink OTFS-NOMA is evaluated. As discussed in Section VI, the NOMA users have two choices for their transmission rates, namely adaptive and fixed rate transmission. The use of adaptive rate transmission can ensure that the implementation of NOMA is transparent to U_0, which means that U_0's QoS requirements are strictly guaranteed. Since U_0 achieves the same performance for OTFS-NOMA and OTFS-OMA when adaptive rate transmission is used, we only focus on the NOMA users' performance, where the ergodic rate in (49) is used as the criterion. We note that this ergodic rate is the net performance gain of OTFS-NOMA over OTFS-OMA, which is the reason why the vertical axis in Fig. 4 is labeled 'Ergodic Rate Gain'. When the M users are randomly selected from the K NOMA users, the ergodic rate gain is moderate, e.g., 1.5 bit per channel use (BPCU) at $\rho = 30$ dB. By applying the scheduling strategy proposed in (50), the ergodic rate gain can be significantly improved, e.g., nearly by a factor of two compared to the random case with $K = 16$ and $\rho = 30$ dB.

Fig. 5 focuses on the case with fixed rate transmission, and similar to Fig. 1, the normalized outage sum rate is used as performance criterion in Fig. 5(a). One can observe that with random user scheduling, the sum rate of OTFS-NOMA is similar to that of OTFS-OMA. This is due to the fact that no interference mitigation strategy, such as power or rate allocation, is used for NOMA uplink transmission, which means that U_0 and the NOMA users cause strong interference to each other and SIC failure may happen frequently. By applying the user scheduling strategy proposed in (50), the channel conditions of the scheduled users become quite different, which facilitates the implementation of SIC. This benefit of user scheduling can be clearly observed in Fig. 5(a), where NOMA achieves a significant gain over OMA although advanced power or rate allocation strategies are not used. Fig. 5(a) also shows that the difference between the performance of FD-LE and FD-DFE is insignificant for the uplink case. This is due to the fact that the outage events during the first stage of SIC dominate the outage performance, and they are not affected by whether FD-LE or FD-DFE is employed. Another important observation from Fig. 5(a) is that the maximal sum rate $R_0 + R_i$ cannot be realized, even at high SNR. The reason for this behaviour is the existence of the error floor for the NOMA users' outage probabilities, as shown in Fig. 5(b). The analytical results provided in Section V-B show that increasing K can reduce the error floor, which is confirmed by Fig. 5(b).

VIII. Conclusion

In this paper, we have proposed OTFS-NOMA uplink and downlink transmission schemes, where users with different mobility profiles are grouped together for the implementation of NOMA. The analytical results developed in the paper demonstrate that both the high-mobility and the low-mobility users benefit from the application of OTFS-NOMA. In particular, the use of NOMA enables the spreading of the signals of a high-mobility user over a large amount of time-frequency resources, which enhances the OTFS resolution and improves the detection reliability. In addition, OTFS-NOMA ensures that the low-mobility users have access to the bandwidth resources which would be solely occupied by the high-mobility users in OTFS-OMA. Hence, OTFS-NOMA improves the spectral efficiency and reduces latency. An interesting topic for future works is studying the impact of non-zero fractional delays and fractional Doppler shifts on the performance of the developed OTFS-NOMA protocol. Furthermore, in this paper, the users' channel gains (the taps of the delay-Doppler impulse response) have been assumed to be Gaussian distributed, and an important direction for future research is to investigate the impact of other types of channel distributions on the performance of OTFS-NOMA. Moreover, the combination of emerging spectrally efficient 5G solutions, such as 5G New Radio Bandwidth Part (5G-NR-BWP) [39], [40] and software-controlled metasurfaces [41], with OTFS-NOMA is also a promising topic for future research.

APPENDIX A
PROOF FOR PROPOSITION 1

Intuitively, the use of $\mathbf{F}_N \otimes \mathbf{F}_M^H$ is analogous to the application of the ISFFT which transforms signals from the delay-Doppler plane to the time-frequency plane, where inter-symbol interference is removed, i.e., the user's channel matrix is diagonalized. The following proof confirms this intuition and reveals how the diagonalized channel matrix is related to the original block circulant matrix. We first apply $\mathbf{F}_N \otimes \mathbf{I}_M$ to $\mathbf{y}_0$, which yields the following:

$$\begin{aligned}&(\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{y}_0\\ &= (\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{H}_0\left(\gamma_0\mathbf{x}_0 + \sum_{q=1}^{M}\gamma_q\mathbf{x}_q\right) + (\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{z}_0\\ &= \operatorname{diag}\left\{\sum_{n=0}^{N-1}\mathbf{A}_{0,n}e^{-j\frac{2\pi ln}{N}}, 0 \le l \le N-1\right\}(\mathbf{F}_N \otimes \mathbf{I}_M)\\ &\quad\times\left(\gamma_0\mathbf{x}_0 + \sum_{q=1}^{M}\gamma_q\mathbf{x}_q\right) + (\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{z}_0, \qquad (61)\end{aligned}$$

where $\operatorname{diag}\{\mathbf{B}_1, \cdots, \mathbf{B}_N\}$ denotes a block-diagonal matrix with $\mathbf{B}_n$, $1 \le n \le N$, on its main diagonal. Note that $\sum_{n=0}^{N-1}\mathbf{A}_{0,n}e^{-j\frac{2\pi ln}{N}}$, $0 \le l \le N-1$, is a sum of N $M \times M$ circulant matrices, each of which can be further diagonalized by $\mathbf{F}_M$. Therefore, we can apply $\mathbf{I}_N \otimes \mathbf{F}_M^H$ to $(\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{y}_0$, which yields the following:

$$\begin{aligned}&(\mathbf{I}_N \otimes \mathbf{F}_M^H)(\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{y}_0\\ &\quad= \operatorname{diag}\left\{\sum_{n=0}^{N-1}\mathbf{\Lambda}_{0,n}e^{-j\frac{2\pi ln}{N}}, 0 \le l \le N-1\right\}\\ &\qquad\times(\mathbf{F}_N \otimes \mathbf{I}_M)(\mathbf{I}_N \otimes \mathbf{F}_M^H)\left(\gamma_0\mathbf{x}_0 + \sum_{q=1}^{M}\gamma_q\mathbf{x}_q\right)\\ &\qquad+ (\mathbf{I}_N \otimes \mathbf{F}_M^H)(\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{z}_0, \qquad (62)\end{aligned}$$

where $\mathbf{\Lambda}_{0,n}$ is a diagonal matrix, $\mathbf{\Lambda}_{0,n} = \operatorname{diag}\left\{\sum_{m=0}^{M-1}a_{0,n}^{m,1}e^{j\frac{2\pi tm}{M}}, 0 \le t \le M-1\right\}$, and $a_{0,n}^{m,1}$ is the element located in the m-th row and first column of $\mathbf{A}_{0,n}$.

By applying a property of the Kronecker product, $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$, the received signals can be simplified as follows:

$$\begin{aligned}&(\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{y}_0\\ &\quad= \underbrace{\operatorname{diag}\left\{\sum_{n=0}^{N-1}\mathbf{\Lambda}_{0,n}e^{-j\frac{2\pi ln}{N}}, 0 \le l \le N-1\right\}}_{\mathbf{D_0}}(\mathbf{F}_N \otimes \mathbf{F}_M^H)\\ &\qquad\times\left(\gamma_0\mathbf{x}_0 + \sum_{q=1}^{M}\gamma_q\mathbf{x}_q\right) + (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{z}_0, \qquad (63)\end{aligned}$$

where the $(kM + l + 1)$-th element on the main diagonal of $\mathbf{D}_0$ is $D_0^{k,l}$ as defined in the proposition. The proof for the proposition is complete.

APPENDIX B
PROOF FOR LEMMA 1

In order to facilitate the SINR analysis, the system model in (18) is further simplified. Define $\tilde{X}[n,m] = \sum_{i=1}^{M}X_i[n,m]$. With the mapping scheme used in (6), the NOMA users' signals are interleaved and orthogonally placed in the time-frequency plane, i.e., $\tilde{X}[n,m]$ is simply U_{m+1}'s n-th signal, $x_{m+1}(n)$. Denote the outcome of the SFFT of $\tilde{X}[n,m]$ by $\tilde{x}[k,l]$, which yields the following transform:

$$\tilde{x}[k,l] = \frac{1}{\sqrt{NM}}\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}\tilde{X}[n,m]e^{-j2\pi\left(\frac{nk}{N}-\frac{ml}{M}\right)}. \qquad (64)$$

Denote the $NM \times 1$ vector collecting the $\tilde{x}[k,l]$ by $\tilde{\mathbf{x}}$ and the $NM \times 1$ vector collecting the $\tilde{X}[n,m]$ by $\check{\mathbf{x}}$, which means that (64) can be rewritten as follows:

$$\tilde{\mathbf{x}} = (\mathbf{F}_N \otimes \mathbf{F}_M^H)\check{\mathbf{x}}. \qquad (65)$$

Therefore, the model for the received signals in (18) can be re-written as follows:

$$\begin{aligned}\check{\mathbf{y}}_0 &= \gamma_0\mathbf{x}_0 + \gamma_1\tilde{\mathbf{x}} + \left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)^{-1}\mathbf{D}_0^{-1}\tilde{\mathbf{z}}_i\\ &= \gamma_0\mathbf{x}_0 + \underbrace{\gamma_1(\mathbf{F}_N \otimes \mathbf{F}_M^H)\check{\mathbf{x}} + \left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)^{-1}\mathbf{D}_0^{-1}\tilde{\mathbf{z}}_0}_{\text{Interference and noise terms}}, \qquad (66)\end{aligned}$$

where we have used the assumption that $\gamma_i = \gamma_1$, for $1 \le i \le N$. Note that the power of the information-bearing signals is simply $\gamma_0^2\rho$, and therefore, the key step to obtain the SINR is to find the covariance matrix of the interference-plus-noise term.

We first show that $\tilde{\mathbf{z}}_0 \triangleq (\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathbf{z}_0$ is still a complex Gaussian vector, i.e., $\tilde{\mathbf{z}}_i \sim CN(0, \mathbf{I}_{NM})$. Recall that $\mathbf{z}_0$ contains NM i.i.d. complex Gaussian random variables. Furthermore, $\mathbf{F}_N \otimes \mathbf{F}_M^H$ is a unitary matrix as shown in the following:

$$\begin{aligned}(\mathbf{F}_N \otimes \mathbf{F}_M^H)(\mathbf{F}_N \otimes \mathbf{F}_M^H)^H &\overset{(a)}{=} (\mathbf{F}_N \otimes \mathbf{F}_M^H)(\mathbf{F}_N^H \otimes \mathbf{F}_M)\\ &\overset{(b)}{=} (\mathbf{F}_N\mathbf{F}_N^H) \otimes (\mathbf{F}_M^H\mathbf{F}_M)\\ &= \mathbf{I}_{NM}, \qquad (67)\end{aligned}$$

where step (a) follows from the fact that $(\mathbf{A} \otimes \mathbf{B})^H = \mathbf{A}^H \otimes \mathbf{B}^H$ and step (b) follows from the fact that

$(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{AC})\otimes(\mathbf{BD})$. Therefore, $(\mathbf{F}_N\otimes\mathbf{F}_M^H)\mathbf{z}_0 \sim CN(0,\mathbf{I}_{NM})$ given the fact that $\mathbf{z}_0 \sim CN(0,\mathbf{I}_{NM})$ and a unitary transformation of a Gaussian vector is still a Gaussian vector.

Therefore, the covariance matrix of the interference-plus-noise term is given by

$$\begin{aligned}
&\mathbf{C}_{\text{cov}} \\
&= \gamma_1^2\mathcal{E}\left\{(\mathbf{F}_N\otimes\mathbf{F}_M^H)\check{\mathbf{x}}\check{\mathbf{x}}^H\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)^H\right\} \\
&\quad + \mathcal{E}\left\{\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)^{-1}\mathbf{D}_0^{-1}\tilde{\mathbf{z}}_0\tilde{\mathbf{z}}_0^H\mathbf{D}_0^{-H}\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)^{-H}\right\}.
\end{aligned} \tag{68}$$

Recall that the $(nM+m+1)$-th element of $\check{\mathbf{x}}$ is $\tilde{X}[n,m]$ which is equal to $x_{m+1}(n)$. Therefore, the covariance matrix can be further simplified as follows:

$$\begin{aligned}
\mathbf{C}_{\text{cov}} &= \gamma_1^2\rho(\mathbf{F}_N\otimes\mathbf{F}_M^H)\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)^H \\
&\quad + \left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)^{-1}\mathbf{D}_0^{-1}\mathbf{D}_0^{-H}\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)^{-H} \\
&= \gamma_1^2\rho\mathbf{I}_{MN} + \left(\mathbf{F}_N^H\otimes\mathbf{F}_M\right)\mathbf{D}_0^{-1}\mathbf{D}_0^{-H}\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right),
\end{aligned} \tag{69}$$

where the noise power is assumed to be normalized.

Following the same steps as in the proof of Proposition 1, we learn that, by construction, $\left(\mathbf{F}_N^H\otimes\mathbf{F}_M\right)\mathbf{D}_0^{-1}\mathbf{D}_0^{-H}\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)$ is also a block-circulant matrix, which means that the elements on the main diagonal of $\left(\mathbf{F}_N^H\otimes\mathbf{F}_M\right)\mathbf{D}_0^{-1}\mathbf{D}_0^{-H}\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)$ are identical. Without loss of generality, denote the diagonal elements of $\left(\mathbf{F}_N^H\otimes\mathbf{F}_M\right)\mathbf{D}_0^{-1}\mathbf{D}_0^{-H}\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)$ by ϕ. Therefore, ϕ can be found by using the trace of the matrix as follows:

$$\begin{aligned}
\phi &= \frac{1}{NM}\text{Tr}\left\{\left(\mathbf{F}_N^H\otimes\mathbf{F}_M\right)\mathbf{D}_0^{-1}\mathbf{D}_0^{-H}\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)\right\} \\
&= \frac{1}{NM}\text{Tr}\left\{\left(\mathbf{F}_N\otimes\mathbf{F}_M^H\right)\left(\mathbf{F}_N^H\otimes\mathbf{F}_M\right)\mathbf{D}_0^{-1}\mathbf{D}_0^{-H}\right\} \\
&= \frac{1}{NM}\text{Tr}\left\{\mathbf{D}_0^{-1}\mathbf{D}_0^{-H}\right\} = \frac{1}{NM}\sum_{k=0}^{N-1}\sum_{l=0}^{M-1}|D_0^{k,l}|^{-2}.
\end{aligned} \tag{70}$$

Therefore, the SINR for detection of $x_0[k,l]$ is given by

$$\text{SINR}_{0,kl}^{LE} = \frac{\rho\gamma_0^2}{\rho\gamma_1^2+\phi}, \tag{71}$$

and the proof is complete.

Appendix C
Proof for Lemma 2

The lemma is proved by first developing upper and lower bounds on the outage probability, and then showing that both bounds have the same diversity order.

An upper bound on $\text{SINR}_{0,kl}$ is given by

$$\begin{aligned}
\text{SINR}_{0,kl} &= \frac{\rho\gamma_0^2}{\rho\gamma_1^2+\frac{1}{NM}\sum_{\tilde{k}=0}^{N-1}\sum_{\tilde{l}=0}^{M-1}|D_0^{\tilde{k},\tilde{l}}|^{-2}} \\
&\leq \frac{\rho\gamma_0^2}{\rho\gamma_1^2+\frac{1}{NM}|D_0^{0,0}|^{-2}}.
\end{aligned} \tag{72}$$

Therefore, the outage probability, denoted by $\text{P}_{0,kl}$, can be lower bounded as follows:

$$\begin{aligned}
\text{P}_{0,kl} &\geq \text{P}\left(\frac{\rho\gamma_0^2}{\rho\gamma_1^2+\frac{1}{NM}|D_0^{0,0}|^{-2}} < \epsilon_0\right) \\
&= \text{P}\left(|D_0^{0,0}|^2 < \frac{\epsilon_0}{NM\rho(\gamma_0^2-\gamma_1^2\epsilon_0)}\right),
\end{aligned} \tag{73}$$

where we assume that $\gamma_0^2 > \gamma_1^2\epsilon_0$. Otherwise, the outage probability is always one.

To evaluate the lower bound on the outage probability, the distribution of $D_0^{u,v}$ is required. Recall from (16) that $D_0^{u,v}$ is the $((v-1)M+u)$-th main diagonal element of $\mathbf{D}_0$ and can be expressed as follows:

$$D_0^{u,v} = \sum_{n=0}^{N-1}\sum_{m=0}^{M-1}a_{0,n}^{m,1}e^{j2\pi\frac{um}{M}}e^{-j2\pi\frac{vn}{N}}, \tag{74}$$

which is the ISFFT of $a_{0,n}^{m,1}$. Therefore, we have the following property:

$$\tilde{\mathbf{D}}_0 = \sqrt{NM}\mathbf{F}_M^H\mathbf{A}_0\mathbf{F}_N, \tag{75}$$

where the element in the u-th row and the v-th column of $\tilde{\mathbf{D}}_0$ is $D_0^{u,v}$ and the element in the m-th row and the n-th column of $\mathbf{A}_0$ is $a_{0,n}^{m,1}$.

The matrix-based expression shown in (75) can be vectorized as follows:

$$\begin{aligned}
\text{Diag}(\mathbf{D}_0) &= \text{vec}(\tilde{\mathbf{D}}_0) = \sqrt{NM}\text{vec}(\mathbf{F}_M^H\mathbf{A}_0\mathbf{F}_N) \\
&= \sqrt{NM}(\mathbf{F}_N\otimes\mathbf{F}_M^H)\text{vec}(\mathbf{A}_0),
\end{aligned} \tag{76}$$

where $\text{Diag}(\mathbf{A})$ denotes a vector collecting all elements on the main diagonal of $\mathbf{A}$ and we use the facts that $(\mathbf{C}^T\otimes\mathbf{A})\text{vec}(\mathbf{B}) = \text{vec}(\mathbf{D})$ if $\mathbf{ABC}=\mathbf{D}$, and $\mathbf{F}_N^T = \mathbf{F}_N$.

We note that $\text{vec}(\mathbf{A}_0)$ contains only (P_0+1) non-zero elements, where the remaining elements are zero. Therefore, each element on the main diagonal of $\mathbf{D}_0$ is a superposition of (P_0+1) i.i.d. random variables, $h_{i,p} \sim CN\left(0,\frac{1}{P_0+1}\right)$. We further note that the coefficients for the superposition are complex exponential constants, i.e., the magnitude of each coefficient is one. Therefore, each element on the main diagonal of $\mathbf{D}_0$ is still complex Gaussian distributed, i.e., $D_0^{u,v} \sim CN(0,1)$, which means that the lower bound on the outage probability shown in (73) can be expressed as follows:

$$\text{P}_{0,kl} \geq 1 - e^{-\frac{\epsilon_0}{NM\rho(\gamma_0^2-\gamma_1^2\epsilon_0)}} \doteq \frac{1}{\rho}. \tag{77}$$

On the other hand, an upper bound on the outage probability is given by

$$\text{P}_{0,kl} \leq \text{P}\left(\frac{\rho\gamma_0^2}{\rho\gamma_1^2+\frac{1}{NM}\sum_{\tilde{k}=0}^{N-1}\sum_{\tilde{l}=0}^{M-1}|D_0^{\min}|^{-2}} < \epsilon_0\right), \tag{78}$$

where $|D_0^{\min}| = \min\{|D_0^{k,l}|, \forall l \in \{0,\cdots,M-1\}, k \in \{0,\cdots,N-1\}\}$.

Therefore, the outage probability can be upper bounded as follows:

$$\text{P}_{0,kl} \leq \text{P}\left(|D_0^{\min}|^2 < \frac{\epsilon_0}{\rho(\gamma_0^2-\gamma_1^2\epsilon_0)}\right). \tag{79}$$

It is important to point out that the $|D_0^{k,l}|^2$, $l \in \{0, \cdots, M-1\}, k \in \{0, \cdots, N-1\}$, are identically but not independently distributed. This correlation property is shown as follows. The covariance matrix of the effective channel gains, i.e., the elements on the main diagonal of $\mathbf{D}_0$, is given by

$$\begin{aligned}
&\mathcal{E}\left\{\mathrm{Diag}(\mathbf{D}_0)\mathrm{Diag}(\mathbf{D}_0)^{\mathrm{H}}\right\} \\
&= NM\mathcal{E}\left\{(\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathrm{vec}(\mathbf{A}_0)\mathrm{vec}(\mathbf{A}_0)^H(\mathbf{F}_N \otimes \mathbf{F}_M^H)^H\right\} \\
&= NM(\mathbf{F}_N \otimes \mathbf{F}_M^H)\mathcal{E}\left\{\mathrm{vec}(\mathbf{A}_0)\mathrm{vec}(\mathbf{A}_0)^H\right\}(\mathbf{F}_N \otimes \mathbf{F}_M^H)^H.
\end{aligned} \tag{80}$$

Because the channel gains, $h_{0,p}$, are i.i.d., $\mathcal{E}\left\{\mathrm{vec}(\mathbf{A}_0)\mathrm{vec}(\mathbf{A}_0)^H\right\}$ is a diagonal matrix, where only (P_0+1) of its main diagonal elements are non-zero. Following the same steps as in the proof for Proposition 1, one can show that the product of $(\mathbf{F}_N \otimes \mathbf{F}_M^H)$, a diagonal matrix, and $(\mathbf{F}_N \otimes \mathbf{F}_M^H)^H$ yields a block circulant matrix, which means that $\mathcal{E}\left\{\mathrm{Diag}(\mathbf{D}_0)\mathrm{Diag}(\mathbf{D}_0)^{\mathrm{H}}\right\}$ is a block-circulant matrix, not a diagonal matrix. Therefore, the $|D_0^{k,l}|^2$, $l \in \{0, \cdots, M-1\}, k \in \{0, \cdots, N-1\}$, are correlated, and not independent.

Although the $|D_0^{k,l}|^2$ are not independent, an upper bound on $\mathrm{P}_{0,kl}$ can be still found as follows:

$$\begin{aligned}
\mathrm{P}_{0,kl} &\leq \mathrm{P}\left(|D_0^{\min}|^2 < \frac{\epsilon_0}{\rho(\gamma_0^2 - \gamma_1^2\epsilon_0)}\right) \\
&\leq \sum_{k=0}^{N-1}\sum_{l=0}^{M-1} \mathrm{P}\left(|D_0^{k,l}|^2 < \frac{\epsilon_0}{\rho(\gamma_0^2 - \gamma_1^2\epsilon_0)}\right) \\
&\leq MN\mathrm{P}\left(|D_0^{0,0}|^2 < \frac{\epsilon_0}{\rho(\gamma_0^2 - \gamma_1^2\epsilon_0)}\right) \\
&= MN\left(1 - e^{-\frac{\epsilon_0}{\rho(\gamma_0^2 - \gamma_1^2\epsilon_0)}}\right) \doteq \frac{1}{\rho}.
\end{aligned} \tag{81}$$

Since both the upper and lower bounds on the outage probability have the same diversity order, the proof of the lemma is complete.

References

[1] Z. Ding *et al.*, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[2] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 176–183, Oct. 2017.

[3] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[4] H. Sari, F. Vanhaverbeke, and M. Moeneclaey, "Multiple access using two sets of orthogonal signal waveforms," *IEEE Commun. Lett.*, vol. 4, no. 1, pp. 4–6, Jan. 2000.

[5] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE 24th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, London, U.K., Sep. 2013, pp. 611–615.

[6] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[7] 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.

[8] A. Brighente and S. Tomasin, "Power allocation for non-orthogonal millimeter wave systems with mixed traffic," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 432–443, Jan. 2019.

[9] Y. Wu, L. P. Qian, K. Ni, C. Zhang, and X. Shen, "Delay-minimization nonorthogonal multiple access enabled multi-user mobile edge computation offloading," *IEEE J. Sel. Topics Signal Process*, vol. 13, no. 3, pp. 392–407, Jun. 2019.

[10] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, 2017.

[11] Y. Zhou, V. W. S. Wong, and R. Schober, "Coverage and rate analysis of millimeter wave NOMA networks with beam misalignment," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8211–8227, Dec. 2018.

[12] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 629–633, May 2016.

[13] R. Chopra, C. R. Murthy, H. A. Suraweera, and E. G. Larsson, "Analysis of nonorthogonal training in massive MIMO under channel aging with SIC receivers," *IEEE Signal Process. Lett.*, vol. 26, no. 2, pp. 282–286, Feb. 2019.

[14] A. Maatouk, E. Çalışkan, M. Koca, M. Assaad, G. Gui, and H. Sari, "Frequency-domain NOMA with two sets of orthogonal signal waveforms," *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 906–909, May 2018.

[15] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[16] L. Yin, W. O. Popoola, X. Wu, and H. Haas, "Performance evaluation of non-orthogonal multiple access in visible light communication," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 5162–5175, Dec. 2016.

[17] H. Marshoud, V. M. Kapinas, G. K. Karagiannidis, and S. Muhaidat, "Non-orthogonal multiple access for visible light communications," *IEEE Photon. Technol. Lett.*, vol. 28, no. 1, pp. 51–54, Jan. 1, 2016.

[18] Z. Ding, P. Fan, and H. V. Poor, "Impact of non-orthogonal multiple access on the offloading of mobile edge computing," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 375–390, Jan. 2019.

[19] *Study on Downlink Multiuser Superposition Transmission for LTE*, document TR 36.859, 3GPP, Mar. 2015.

[20] *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation (Release 15)*, document TS 36.211, 3GPP, Jan. 2019.

[21] *Study on Non-Orthogonal Multiple Access (NOMA) for NR (Release 16)*, document TR 38.812, 3GPP, Dec. 2018.

[22] B. Di, L. Song, Y. Li, and Z. Han, "V2X meets NOMA: Non-orthogonal multiple access for 5G-enabled vehicular networks," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 14–21, Dec. 2017.

[23] Y. Chen, L. Wang, Y. Ai, B. Jiao, and L. Hanzo, "Performance analysis of NOMA-SM in vehicle-to-vehicle massive MIMO channels," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2653–2666, Dec. 2017.

[24] R. Hadani and A. Monk, "OTFS: A new generation of modulation addressing the challenges of 5G," 2018, *arXiv:1802.02623*. [Online]. Available: https://arxiv.org/abs/1802.02623

[25] R. Hadani *et al.*, "Orthogonal time frequency space modulation," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.

[26] K. R. Murali and A. Chockalingam, "On OTFS modulation for high-Doppler fading channels," in *Proc. Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, Feb. 2018, pp. 1–10.

[27] P. Raviteja, Y. Hong, E. Viterbo, and E. Biglieri, "Practical pulse-shaping waveforms for reduced-cyclic-prefix OTFS," *IEEE Trans. Veh. Tech.*, vol. 68, no. 1, pp. 957–961, Jan. 2019.

[28] P. Raviteja, K. T. Phan, Y. Hong, and E. Viterbo, "Interference cancellation and iterative detection for orthogonal time frequency space modulation," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6501–6515, Oct. 2018.

[29] G. D. Surabhi, R. M. Augustine, and A. Chockalingam, "On the diversity of uncoded OTFS modulation in doubly-dispersive channels," 2018, *arXiv:1808.07747*. [Online]. Available: https://arxiv.org/abs/1808.07747

[30] V. Khammammetti and S. K. Mohammed, "OTFS-based multiple-access in high Doppler and delay spread wireless channels," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 528–531, Apr. 2019.

[31] A. RezazadehReyhani, A. Farhang, M. Ji, R. R. Chen, and B. Farhang-Boroujeny, "Analysis of discrete-time MIMO OFDM-based orthogonal time frequency space modulation," in *Proc. IEEE Int. Conf. Communicat. (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–6.

[32] M. K. Ramachandran and A. Chockalingam, "MIMO-OTFS in high-Doppler fading channels: Signal detection and channel estimation," in *Proc. IEEE GLOBECOM*, Kansas City, MO, USA, Dec. 2018, pp. 206–212.

[33] P. Raviteja, E. Viterbo, and Y. Hong, "OTFS performance on static multipath channels," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 745–748, Jun. 2019.

[34] D. Falconer, S. L. Ariyavisitakul, A. Benyamin-Seeyar, and B. Eidson, "Frequency domain equalization for single-carrier broadband wireless systems," *IEEE Commun. Mag.*, vol. 40, no. 4, pp. 58–66, Apr. 2002.

[35] J. Louveaux, L. Vandendorpe, and T. Sartenaer, "Cyclic prefixed single carrier and multicarrier transmission: Bit rate comparison," *IEEE Commun. Lett.*, vol. 7, no. 4, pp. 180–182, Apr. 2003.

[36] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.

[37] B. Devillers, "Cyclic prefixed block transmission for wireless communications: Performance analysis and optimization," Ph.D. dissertation, Université catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium, 2009. AQ:5

[38] B. Devillers, J. Louveaux, and L. Vandendorpe, "About the diversity in cyclic prefixed single-carrier systems," *Phys. Commun.*, vol. 1, no. 4, pp. 266–276, Dec. 2008.

[39] J. Jeon, "NR wide bandwidth operations," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 42–46, Mar. 2018.

[40] C. Sexton, N. Marchetti, and L. A. DaSilva, "Customization and trade-offs in 5G RAN slicing," *IEEE Commun. Mag.*, vol. 57, no. 4, pp. 116–122, Apr. 2019.

[41] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "A new wireless communication paradigm through software-controlled metasurfaces," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 162–169, Sep. 2018.

AQ:6

Zhiguo Ding (S'03–M'05–SM'15) received the B.Eng. degree in electrical engineering from the Beijing University of Posts and Telecommunications in 2000 and the Ph.D. degree in electrical engineering from Imperial College London in 2005.

From July 2005 to April 2018, he was with Queen's University Belfast, Imperial College, Newcastle University, and Lancaster University. Since April 2018, he has been a Professor of communications with The University of Manchester. From October 2012 to September 2018, he was an Academic Visitor with Princeton University. His research interests are 5G networks, game theory, cooperative and energy harvesting networks, and statistical signal processing. He was a recipient of the Best Paper Award at IET ICWMC-2009 and IEEE WCSP-2014, the EU Marie Curie Fellowship (2012–2014), the Top IEEE TVT Editor 2017, the 2018 IEEE Communication Society Heinrich Hertz Award, the 2018 IEEE Vehicular Technology Society Jack Neubauer Memorial Award, and the 2018 IEEE Signal Processing Society Best Signal Processing Letter Award. He was an Editor of IEEE WIRELESS COMMUNICATIONS LETTERS and IEEE COMMUNICATIONS LETTERS from 2013 to 2016. He has been serving as an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and *Journal of Wireless Communications and Mobile Computing*.

Robert Schober (M'01–SM'08–F'10) received the Diploma (Univ.) and Ph.D. degrees in electrical engineering from the Friedrich-Alexander University of Erlangen-Nuremberg (FAU), Germany, in 1997 and 2000, respectively.

From 2002 to 2011, he was a Professor and the Canada Research Chair with The University of British Columbia (UBC), Vancouver, Canada. Since January 2012, he has been an Alexander von Humboldt Professor and the Chair for Digital Communication with FAU. His research interests fall into the broad areas of communication theory, wireless communications, and statistical signal processing. He is also a fellow of the Canadian Academy of Engineering and the Engineering Institute of Canada. He was a recipient of several awards for his work, including the 2002 Heinz Maier-Leibnitz Award of the German Science Foundation (DFG), the 2004 Innovations Award of the Vodafone Foundation for Research in Mobile Communications, the 2006 UBC Killam Research Prize, the 2007 Wilhelm Friedrich Bessel Research Award of the Alexander von Humboldt Foundation, the 2008 Charles McDowell Award for Excellence in Research from UBC, the 2011 Alexander von Humboldt Professorship, the 2012 NSERC E.W.R. Steacie Fellowship, and the 2017 Wireless Communications Recognition Award by the IEEE Wireless Communications Technical Committee. He was listed as a 2017 Highly Cited Researcher by the Web of Science. He is also the Chair of the Steering Committee of IEEE TRANSACTIONS ON MOLECULAR, BIOLOGICAL AND MULTI-SCALE COMMUNICATION, a member of the Editorial Board of PROCEEDINGS OF THE IEEE, a Member-at-Large of the Board of Governors of ComSoc, and the ComSoc Director of journals. He is also a Distinguished Lecturer of the IEEE Communications Society (ComSoc). From 2012 to 2015, he served as the Editor-in-Chief of IEEE TRANSACTIONS ON COMMUNICATIONS.

Pingzhi Fan (M'93–SM'99–F'15) received the M.Sc. degree in computer science from Southwest Jiaotong University, China, in 1987, and the Ph.D. degree in electronic engineering from Hull University, U.K., in 1994.

He was the Chief Scientist of the National 973 Research Project (MoST) from 2012 to 2016. He is currently a Professor and the Director of the Institute of Mobile Communications, Southwest Jiaotong University. He has been a Visiting Professor with Leeds University, U.K., since 1997, and has been a Guest Professor with Shanghai Jiaotong University since 1999. He has over 280 research papers published in various international journals and eight books (including edited). He is the inventor of 22 granted patents. His research interests include vehicular communications, wireless networks for big data, and signal design and coding. He is also a fellow of IET, CIE, and CIC. He was a recipient of the U.K. ORS Award in 1992 and the Outstanding Young Scientist Award (NSFC) in 1998. He has served as the general chair or the TPC chair of a number of international conferences. He is also the Founding Chair of IEEE VTS BJ Chapter, IEEE ComSoc CD Chapter, and IEEE Chengdu Section. He is also the guest editor or editorial member of several international journals. He has also served as the Board Member of IEEE Region 10, IET (IEE) Council, and IET Asia–Pacific Region. He is also an IEEE VTS Distinguished Lecturer (2015–2019).

H. Vincent Poor (M'77–SM'82–F'87) received the Ph.D. degree in EECS from Princeton University in 1977.

From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana–Champaign. Since 1990, he has been on the faculty at Princeton University, where he is currently the Michael Henry Strater University Professor of electrical engineering. From 2006 to 2016, he served as the Dean of the School of Engineering and Applied Science, Princeton University. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory and signal processing, and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the recent book *Multiple Access Techniques for 5G Wireless Networks and Beyond.* (Springer, 2019). He is also a member of the National Academy of Engineering and the National Academy of Sciences. He is also a Foreign Member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He was a recipient of the Marconi and Armstrong Awards of the IEEE Communications Society in 2007 and 2009, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal, the 2019 ASEE Benjamin Garver Lamme Award, the D.Sc. *(honoris causa)* from Syracuse University awarded in 2017, and the D.Eng. *(honoris causa)* from the University of Waterloo awarded in 2019.