Low-Complexity Reliability-Based Equalization and Detection for OTFS-NOMA

Stephen McWade, *Member, IEEE,* Arman Farhang, *Senior Member, IEEE,* and Mark F. Flanagan, *Senior Member, IEEE*

Abstract—**Orthogonal time frequency space (OTFS) modulation has recently emerged as a potential 6G candidate waveform which provides improved performance in high-mobility scenarios. In this paper we investigate the combination of OTFS with non-orthogonal multiple access (NOMA). Existing equalization and detection methods for OTFS-NOMA, such as minimum-mean-squared error with successive interference cancellation (MMSE-SIC), suffer from poor performance. Additionally, existing iterative methods for single-user OTFS based on low-complexity iterative least-squares solvers are not directly applicable to the NOMA scenario due to the presence of multi-user interference (MUI). Motivated by this, in this paper we propose a low-complexity method for equalization and detection for OTFS-NOMA. The proposed method uses a novel reliability zone (RZ) detection scheme which estimates the reliable symbols of the users and then uses interference cancellation to remove MUI. The thresholds for the RZ detector are optimized in a greedy manner to further improve detection performance. In order to optimize these thresholds, we modify the least squares with QR-factorization (LSQR) algorithm used for channel equalization to compute the the post-equalization mean-squared error (MSE), and track the evolution of this MSE throughout the iterative detection process. Numerical results demonstrate the superiority of the proposed equalization and detection technique to the existing MMSE-SIC benchmark in terms of symbol error rate (SER).**

I. Introduction

The sixth generation (6G) of mobile networks is expected to support communications in high-mobility environments such as high-speed rail, vehicle-to-everything (V2X) and unmanned aerial vehicle (UAV) communications [1]. Orthogonal frequency division multiplexing (OFDM) has been the waveform utilized in the 4th and 5th generation of wireless networks. However, it is well-known that in high-mobility scenarios, OFDM performs poorly due to the Doppler effect [2]. In recent years, a new waveform called orthogonal time frequency space (OTFS) has been proposed to address this drawback of OFDM in time-varying channels. In contrast to OFDM, which transmits data symbols in the time-frequency domain, OTFS places the data symbols in the delay-Doppler domain [3]. OTFS then uses a transformation to spread each information symbol over the whole time-frequency plane. This means that the symbols are all equally affected by the time and frequency selectivity of the channel which converts the time-varying channel to a time-invariant one in the delay-Doppler domain.

A number of OTFS equalization and detection schemes have been proposed in the literature in recent years. The majority of these methods can be categorized into either low-complexity linear equalizers [4]–[6] or non-linear message-passing-based equalizers [7]–[9]. However, such methods assume a scattering environment in which the channel impulse response is sparse in the delay-Doppler domain. Under more realistic channel conditions, the low-complexity linear schemes are no longer applicable as the assumptions they make about the channel no longer hold. Additionally, message-passing-based detectors become prohibitively complex due to the large number of scatterers [10]. An alternative approach was proposed in [10] which utilized a least-squares minimum residual (LSMR) based channel equalizer and a reliability-based dynamic detector. However, the system model in [10] only considers a single-user scenario and it is not applicable to the multi-user scenario that is of interest in this paper.

For a multi-user OTFS system, the multiple access (MA) technique utilized is an important consideration. How best to multiplex users in the delay-Doppler domain is an open question and there have been numerous recent works which propose different methods [11]–[13]. These methods can be broadly categorized into orthogonal multiple access (OMA) or non-orthogonal multiple access (NOMA). In OTFS-OMA, users are multiplexed either in the delay domain or the Doppler domain, and only one user can occupy a given resource block [13]. However, the users suffer from multi-user interference (MUI) due to the Doppler spread, which degrades performance. MUI can be mitigated by inserting guard bands between users, as was done in [11]. However, this use of guard bands leads to a spectral efficiency (SE) loss [12].

An alternative approach is OTFS-NOMA, where the users are allowed to occupy the same resource block and are multiplexed in either the power domain or the code domain. A multi-user detection (MUD) scheme, such as successive interference cancellation (SIC), is then used to detect the user symbols [14]. NOMA is a well-known technique which can provide improved SE over the corresponding OMA system as well as potentially higher connectivity as the number of users supported by a NOMA system is not limited by the number

S. McWade was with the School of Electrical and Electronic Engineering, University College Dublin, Belfield, Dublin 4, D04 V1W8 Ireland. He is now with the Department of Electronic and Electrical Engineering, Trinity College Dublin, Dublin 2, D02 PN40 Ireland (email: smcwade@tcd.ie). A. Farhang is with the Department of Electronic and Electrical Engineering, Trinity College Dublin, Dublin 2, D02 PN40 Ireland (email: arman.farhang@tcd.ie). M. F. Flanagan is with the School of Electrical and Electronic Engineering, University College Dublin, Belfield, Dublin 4, D04 V1W8 Ireland (email:mark.flanagan@ieee.org).

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant number 17/RC-PhD/3479, Grant number 17/US/3445, Grant number 19/FFP/7005(T) and under the US-Ireland R&D Partnership Programme Grant Number SFI/21/US/3757.

of physical resources available. A number of OTFS-NOMA schemes have been proposed in the literature in recent years that use either power-domain [15], [16] or code-domain [17], [18] multiplexing. This paper focuses on power-domain OTFS-NOMA.

With regard to the existing work on power-domain OTFS-NOMA, the authors of [15] considered a single high-mobility OTFS user multiplexed with multiple low-mobility OFDM users. However, this system model is restricted to a single OTFS user and hence cannot accommodate multiple high-mobility users. The authors of [16] addressed this issue and proposed an OTFS-NOMA scheme which utilizes a rectangular pulse shape where multiple users overlap in the delay-Doppler domain and are multiplexed in the power domain. The results presented in [16] show that OTFS-NOMA achieves higher spectral efficiency than the equivalent OTFS-OMA system. However, the system proposed in [16] used minimum-mean-squared-error (MMSE) equalization in combination with SIC for equalization and detection. The problem with this scheme is that direct implementation of MMSE equalization is prohibitively computationally complex and thus impractical for real-world scenarios.

As of yet, to the best of our knowledge, there is no low-complexity equalization and detection method for power-domain OTFS-NOMA. In addition, the low-complexity equalization and detection method of [10] for single-user OTFS is not directly applicable to a NOMA scenario due to the presence of MUI. This paper addresses these gaps in the literature with the following contributions:

- We propose a novel iterative method for equalization and detection of a downlink OTFS-NOMA system which, within each iteration, uses a proposed modified LSQR (mLSQR) algorithm to equalize the channel, an RZ detector to detect reliable symbols from both users, and interference cancellation to improve detection on subsequent iterations.
- Our proposed modified LSQR algorithm, in addition to equalizing the channel, also computes the post-equalization MSE of the users' symbols, in contrast to the conventional LSQR algorithm. We derive an exact closed-form expression for this MSE as well as a low-complexity approximation which capitalizes on the properties of the delay-Doppler channel in OTFS systems.
- We use a novel, greedy approach for optimizing the RZ thresholds within each iteration. This is in contrast to other RZ schemes which use heuristic thresholds [10], [19]. Our method works by tracking the post-equalization MSE after interference cancellation and optimizing the RZ thresholds in each iteration to minimize the MSE.

Additionally, we present numerical results which compare the SER performance of the proposed equalization and detection method with the existing MMSE-SIC benchmark [16]. We also compare the performance of our optimized RZ threshold design to a pre-determined threshold design. The presented results demonstrate the superiority of our proposed method, especially for the NOMA user with the smaller power allocation. A preliminary version of this work was described in [20], which showed the advantage of this general approach but did not include the derivation of the post-equalization MSE (or its low-complexity approximation), and also did not show how this MSE could be utilized to optimize the RZ thresholds in each iteration.

The rest of this paper is organized as follows. Section II describes the system model for a 2-user OTFS-NOMA system. In Section III, the proposed equalization and detection algorithm is presented. Section IV describes the modified LSQR algorithm which equalizes the channel and computes the post-equalization MSE. Section V presents the process for optimizing the thresholds of the RZ detector. Section VI presents numerical results. Finally, Section VII concludes the paper.

Notations: Superscripts $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$ denote transpose and Hermitian transpose, respectively. Bold lower-case characters are used to denote vectors and bold upper-case characters are used to denote matrices. The function $\mathrm{vec}\{\mathbf{X}\}$ vectorizes the matrix $\mathbf{X}$ by stacking its columns to form a single column vector, and $\otimes$ represents the Kronecker product. The $p \times p$ identity matrix and $p \times q$ all-zero matrix are denoted by $\mathbf{I}_p$ and $\mathbf{0}_{p\times q}$, respectively.

II. System Model

For ease of exposition, in the following sections we will describe the system model and the proposed detector for the case of a 2-user downlink OTFS-NOMA system; however, note that with appropriate modifications, the proposed method is applicable to any number of users. We consider a downlink OTFS-NOMA system where both users occupy the same delay-Doppler domain resources and are multiplexed in the power domain. For User $i \in \{1, 2\}$, let the $M \times N$ matrix $\mathbf{X}_i$ contain the MN quadrature amplitude modulation (QAM) data symbols placed in the delay-Doppler domain. The elements of $\mathbf{X}_i$ are assumed to be independent and identically distributed (i.i.d.) complex random variables. Additionally, a normalized (unit-energy) square QAM constellation is assumed for each user.

In this work we consider OFDM-based OTFS modulation with rectangular pulse shape. In the first stage of OTFS modulation, the inverse symplectic fast Fourier transform (ISFFT) is used to map the delay-Doppler data symbols in $\mathbf{X}_i$ to the time-frequency domain. The ISFFT can be implemented by performing an M-point DFT operation on each of the columns of $\mathbf{X}_i$ followed by an N-point IDFT operation on each of the rows of $\mathbf{X}_i$. The time-frequency signal matrix of User i is therefore given by

$$\mathbf{D}_i = \mathbf{F}_M \mathbf{X}_i \mathbf{F}_N^{\mathrm{H}}, \tag{1}$$

where $\mathbf{F}_N$ is the N-point unitary discrete Fourier transform (DFT) matrix in which the (l, k) element is $\frac{1}{\sqrt{N}} e^{-j\frac{2\pi}{N}lk}$. Next, cyclic prefix OFDM (CP-OFDM) modulation is used to convert the time-frequency signal to the delay-time domain. The OTFS transmit signal matrix is therefore given by

$$\mathbf{S}_i = \mathbf{A}_{\mathrm{cp}} \mathbf{F}_M^{\mathrm{H}} \mathbf{D}_i, \tag{2}$$

where $\mathbf{A}_{\mathrm{cp}} = [\mathbf{J}_{\mathrm{cp}}, \mathbf{I}_N]$ is the CP addition matrix (here $\mathbf{J}_{\mathrm{cp}}$ is composed of the last N_{cp} rows of $\mathbf{I}_N$). Using (1), the delay-time domain transmit signal can be rewritten as

$$\mathbf{S}_i = \mathbf{A}_{\mathrm{cp}}\mathbf{X}_i\mathbf{F}_N^{\mathrm{H}}, \tag{3}$$

and thus OFDM-based OTFS reduces to an N-point IDFT operation on the rows of $\mathbf{X}_i$ [21]. After parallel to serial conversion, the time-domain symbols for User i can now be written as

$$\mathbf{s}_i = \mathrm{vec}(\mathbf{S}_i) = (\mathbf{F}_N^{\mathrm{H}} \otimes \mathbf{A}_{\mathrm{cp}})\mathbf{x}_i. \tag{4}$$

The users are multiplexed in the power domain and their signals are superimposed before transmission. The superimposed transmit signal is given by

$$\mathbf{s} = \sqrt{\rho_1}\mathbf{s}_1 + \sqrt{\rho_2}\mathbf{s}_2, \tag{5}$$

where ρ_i is the power allocation coefficient for for User i, and $\rho_1+\rho_2 = 1$ (these power allocation coefficients are determined using an appropriate power allocation scheme, such as that used in [16]). Without loss of generality, in this paper, we consider the user indices to be ordered in descending order of their power allocation coefficients, i.e., $\rho_1 > \rho_2$.

After digital to analog conversion, the continuous-time signal $s(t)$ is transmitted through the linear time-varying (LTV) channel. The received signal at the receiver of User $i \in \{1, 2\}$ can be written as

$$r_i(t) = \int\int h_i(\tau, \nu)s(t-\tau)e^{j2\pi\nu(t-\tau)}d\tau d\nu + \omega_i(t) \tag{6}$$

where

$$h_i(\tau, \nu) = \sum_{p=0}^{P_i-1} h_{i,p}\delta(\tau - \tau_{i,p})\delta(\nu - \nu_{i,p}),$$

is the delay-Doppler channel impulse response (CIR) for User i, which consists of P_i channel paths, and $\omega_i(t)$ is the complex AWGN with variance σ_i^2. The parameters $h_{i,p}$, $\tau_{i,p}$ and $\nu_{i,p}$ represent the channel gain, delay and Doppler shift, respectively, associated with path p of User i's channel. The power delay profile (PDP) of the channel of User i is given by $\boldsymbol{\lambda}_i = [\lambda_i(0), \ \ldots \ , \lambda_i(P_i - 1)]$ and is assumed to be normalized such that $\sum_{p=0}^{P_i-1}\lambda_i(p) = 1$. Each channel path gain is modeled as a complex Gaussian random variable with mean zero and and variance $\lambda_i(p)$. Since the PDP is considered to be normalized, the average received SNR of User i is given by $\mathrm{SNR}_i = \frac{\rho_1+\rho_2}{\sigma_i^2}$. We assume perfect knowledge of the User i channel at the receiver of User i, as previously considered in [16].

The received signal is then sampled with sampling period T_{s}. The sampling period is assumed to be short, as is often the case in practical systems, such that the path delays can be considered as integer multiples of the sampling period, i.e., $\tau_{i,p} = l_{i,p}T_{\mathrm{s}}$, where $l_{i,p} \in [0, \ldots, L-1]$. However, the Doppler shifts cannot be considered to be integer multiples of the Doppler spacing and thus, we consider fractional Doppler shifts in this paper. The discrete received signal samples can then be expressed as

$$r_i[n] = \sum_{l=0}^{L-1} h_i[n, l]s[n-l] + \omega_i[n], \tag{7}$$

where $h_i[n, l] = \sum_{p=0}^{P_i-1} h_{i,p}e^{j2\pi\nu_{i,p}(n-l)T_{\mathrm{s}}}\delta(l - l_{i,p})$ is the CIR at time instant n and delay l. The discrete-time received signal can be written in matrix form as

$$\mathbf{r}_i = \mathbf{H}_i\mathbf{s} + \boldsymbol{\omega}_i, \tag{8}$$

where $\boldsymbol{\omega}_i$ is the complex AWGN vector and $\mathbf{H}_i$ is the $MN \times MN$ time-domain channel matrix of User i constructed from the CIRs. The received signal is then demodulated and converted back to the delay-Doppler domain by taking an N-point DFT operation across the time domain samples. Thus, the received signal is given by

$$\mathbf{y}_i = (\mathbf{F}_N \otimes \mathbf{R}_{\mathrm{cp}})\mathbf{r}_i. \tag{9}$$

This can alternatively be written as

$$\mathbf{y}_i = \mathbf{G}_i\mathbf{x}_{\mathrm{sup}} + \mathbf{w}_i. \tag{10}$$

where $\mathbf{G}_i = (\mathbf{F}_N \otimes \mathbf{R}_{\mathrm{cp}})\mathbf{H}_i(\mathbf{F}_N^{\mathrm{H}} \otimes \mathbf{A}_{\mathrm{cp}})$ is the effective channel matrix, $\mathbf{x}_{\mathrm{sup}} = \sqrt{\rho_1}\mathbf{x}_1 + \sqrt{\rho_2}\mathbf{x}_2$ is the superimposed delay-Doppler symbol vector and $\mathbf{w}_i = (\mathbf{F}_N \otimes \mathbf{R}_{\mathrm{cp}})\boldsymbol{\omega}_i$ is the noise vector.

III. Proposed Equalization and Detection Technique

Each user needs to equalize the channel and detect its own symbols at its own receiver. One way to do this is to use MMSE equalization in combination with SIC, as in [16], which we refer to as MMSE-SIC. MMSE equalization operates by pre-multiplying the received vector $\mathbf{y}_i$ in (10) by the MMSE equalization matrix given by

$$\mathbf{W}_{\mathrm{MMSE},i} = \left((\mathbf{G}_i^{\mathrm{H}}\mathbf{G}_i + \sigma_i^2\mathbf{I})^{-1}\right)\mathbf{G}_i^{\mathrm{H}}. \tag{11}$$

More specifically, User 1 uses $\mathbf{W}_{\mathrm{MMSE},1}$ to equalize the channel and then detect its own data symbols while treating the User 2 data symbols as noise. On the other hand, User 2 uses $\mathbf{W}_{\mathrm{MMSE},2}$ to equalize the channel and first detect the User 1 symbols, treating its own symbols as noise. User 2 then removes the User 1 signal from the received signal, uses $\mathbf{W}_{\mathrm{MMSE},2}$ to equalize the channel and then detects its own data symbols [16]. MMSE equalization is impractical for real-world applications due to the $MN \times MN$ matrix inversion in (11), which has a computational complexity of $\mathcal{O}(M^3N^3)$. This is clearly unrealistic for practical applications where M and N can be large. Additionally, while low-complexity implementations of MMSE equalization exist, they assume ideal pulses and a small number of channel scatterers, and thus are not applicable to practical scenarios [10].

The proposed method is inspired by the method proposed in [10] for single-user OTFS which utilized an iterative LSMR-based method with RZ detection and interference cancellation. Note that if the method in [10] is applied directly to OTFS-NOMA with SIC to detect the signals of User 1 and User 2, we can expect poor performance due to the MUI present in the system. Therefore, in the proposed method, we perform SIC at a symbol level rather than a packet level as is done in the MMSE-SIC approach. This allows for the decoding of symbols from both users as soon as they become reliable and also allows for the incorporation of MUI cancellation to

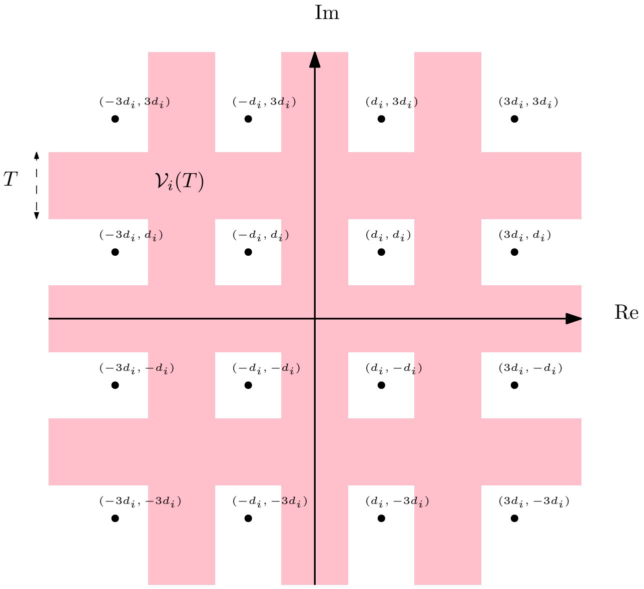

Fig. 1. Illustration of the unreliable zone $\mathcal{V}_i(T)$ in the case where User i employs a 16-QAM constellation.

improve the detection performance. The proposed algorithm uses an iterative process in which the mLSQR algorithm is used to equalize the channel and an RZ detector is used to detect the reliable symbols of both User 1 *and* User 2 within each iteration. Interference cancellation is then used to remove ISI, IDI *and* MUI from the undetected symbols of both users, which improves the detection quality in subsequent iterations. The proposed mLSQR algorithm, which equalizes the channel and also computes the post-equalization MSE, will be explained in detail in Section IV. In the next subsection, we describe the RZ detection process.

A. Reliability zone detector

Here, we first introduce some relevant notation. Each User $i \in \{1,2\}$ uses A_i-ary QAM modulation, where the QAM symbol constellation is defined as

$$\mathcal{A}_i = \{u + vj \ : \ u, v \in \{(2a-1)d_i : \\ a \in \{-\sqrt{A_i}/2+1, \ \dots \ , \sqrt{A_i}/2\}\}\},$$

where d_i is half the distance between adjacent QAM constellation symbols of User i (the value of d_i is chosen so as to ensure a unit-energy constellation $\mathcal{A}_i$). Next, we define the *unreliable zone* with respect to this QAM constellation as

$$\mathcal{V}_i(T) = \{u + vj \ |u, v \in \mathcal{U}_i(T)\}, \tag{12}$$

where

$$\mathcal{U}_i(T) = \bigcup_{a=-\sqrt{A_i}/2+1}^{\sqrt{A_i}/2-1} \mathcal{U}_{i,a}(T). \tag{13}$$

and

$$\mathcal{U}_{i,a}(T) = \{u \ | \ 2ad_i - T/2 < u < 2ad_i + T/2\},$$

where T is a pre-defined threshold which determines the size of the unreliable zone. To demonstrate, the shaded areas in Fig. 1 shows an illustration of the unreliable zone $\mathcal{V}_i(T)$ for a 16-QAM constellation.

Algorithm 1 Proposed Algorithm for symbol detection at User i receiver

```
1: Input: User index $i$, Channel matrix $\mathbf{G}_i$, received symbol
   vector $\mathbf{y}_i$, power allocation fractions $\rho_1$ and $\rho_2$
2: Initialize: $\mathbf{y}^{(1)} = \mathbf{y}_i$, $\hat{\mathbf{x}}_1 = \hat{\mathbf{x}}_2 = \tilde{\mathbf{x}}_{1,\mathrm{q}} = \tilde{\mathbf{x}}_{2,\mathrm{q}} = \mathbf{0}_{MN\times 1}$
3: Define $\mathcal{N} = \{0, \dots, MN-1\}$, $\mathcal{N}_1 = \mathcal{N}_2 = \mathcal{N}$, $\mathcal{D}_1 = \emptyset$
4: for $k$ = 1 to $K$ do
5:     $[\tilde{\mathbf{x}}_{\mathrm{sup}}, \gamma] = \mathrm{mLSQR}(\mathbf{G}_i, \mathbf{y}^{(k)}, \sigma_i^2)$
6:     $\tilde{\mathbf{x}}_1 = ((\tilde{\mathbf{x}}_{\mathrm{sup}}[m])_{m\in\mathcal{N}_1})/\sqrt{\rho_1}$
7:     $\tilde{\mathbf{x}}_2 = ((\tilde{\mathbf{x}}_{\mathrm{sup}}[m])_{m\in\mathcal{N}_2\cap\mathcal{D}_1})/\sqrt{\rho_2}$
8:     if $i = 1$
9:         Select threshold $T_1$ by solving (40)
10:        Select threshold $T_2$ by solving (44)
11:    else if $i = 2$
12:        Select threshold $T_1$ by solving (48)
13:        Select threshold $T_2$ by solving (40)
14:    end if
15:    Update users' reliable symbol index sets via
           $\mathcal{R}_j = \{n \in \mathcal{N}_j : \tilde{\mathbf{x}}_j[n] \notin \mathcal{V}_j(T_j^{(k)})\}, \forall j \in \{1,2\}$
16:    Quantize reliable symbols:
           $\tilde{\mathbf{x}}_{j,\mathrm{q}}[r] = Q_j(\tilde{\mathbf{x}}_j[r]), \ \forall r \in \mathcal{R}_j, \forall j \in \{1,2\}$
17:    Remove interference:
           $\mathbf{y}^{(k+1)} = \mathbf{y}^{(k)} - \mathbf{G}_i(\sqrt{\rho_1}\tilde{\mathbf{x}}_{1,\mathrm{q}} + \sqrt{\rho_2}\tilde{\mathbf{x}}_{2,\mathrm{q}})$
18:    Store detected symbols in output vectors:
           $\hat{\mathbf{x}}_j = \tilde{\mathbf{x}}_{j,\mathrm{q}}[r], \ \forall r \in \mathcal{R}_j, \forall j \in \{1,2\}$
19:    Reset: $\tilde{\mathbf{x}}_{1,\mathrm{q}} = \mathbf{0}$ and $\tilde{\mathbf{x}}_{2,\mathrm{q}} = \mathbf{0}$
20:    Update: $\mathcal{N}_1 = \{n \in \mathcal{N} : \hat{x}_1[n] = 0\}$, $\mathcal{N}_2 = \{n \in \mathcal{N} :$
       $\hat{x}_2[n] = 0\}$, $\mathcal{D}_1 = \{n \in \mathcal{N} : n \notin \mathcal{N}_1\}$
21:    if $\mathcal{N}_i = \emptyset$, break
22: end for
23: Output: $\hat{\mathbf{x}}_i$
```

In the detection process, decisions are made in a symbol-by-symbol manner. If a symbol $x_i[n]$ is outside $\mathcal{V}_i(T_i)$, then it is deemed reliable and can be quantized to the nearest symbol in $\mathcal{A}_i$; the resulting symbol is denoted by $x_{i,q}[n] = Q_i(x[n])$. If $x_i[n]$ is inside $\mathcal{V}_i(T_i)$ then it is deemed unreliable and no quantization takes place. The detected reliable symbols can then be used for interference cancellation.

B. Proposed algorithm

In this subsection, we describe the proposed method for equalization and detection of the OTFS-NOMA signal at the receiver of User $i \in \{1,2\}$. This method is described in Algorithm 1. Each iteration begins on line 5 of Algorithm 1, where the LSQR algorithm is used to equalize the channel and obtain a new estimate, $\tilde{\mathbf{x}}_{\mathrm{sup}}$, of the superimposed transmitted symbol vector via

$$[\tilde{\mathbf{x}}_{\mathrm{sup}}, \gamma_i] = \mathrm{mLSQR}(\mathbf{G}_i, \mathbf{y}^{(k)}, \sigma_i^2). \tag{14}$$

Additionally, our proposed modification to the LSQR algorithm calculates the post-equalization MSE (denoted by γ) over all of the symbols of both users. The exact workings of the mLSQR algorithm and the role of the MSE γ in optimizing the RZ detector's thresholds will be explained in detail in Sections IV and V, respectively. In lines 6 and 7, two sub-vectors are formed from $\tilde{\mathbf{x}}_{\text{sup}}$. The vector $\tilde{\mathbf{x}}_1$ contains the elements of $\tilde{\mathbf{x}}_{\text{sup}}$ whose indices are in $\mathcal{N}_1$, which is the set of undetected User 1 symbols. Since the RZ detector can only make decisions on User 2 symbols once the corresponding User 1 symbols have been detected on a previous iteration, the vector $\tilde{\mathbf{x}}_2$ contains the elements of $\tilde{\mathbf{x}}_{\text{sup}}$ whose indices are in $\mathcal{N}_2$, the set of undetected User 2 symbols, *and* $\mathcal{D}_1$, the set of detected User 1 symbols. In lines 8 – 14, Algorithm 1 selects the thresholds, T_1 and T_2, to be used in the RZ detector. The exact process for selecting the thresholds will be explained in detail in Section V. Decisions are then made on the reliability of the estimated symbols in $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ via the RZ detector in line 15.

In line 16, the reliable symbols are quantized to the nearest QAM symbol and are stored in the empty vectors $\tilde{\mathbf{x}}_{1,\text{q}}$ and $\tilde{\mathbf{x}}_{2,\text{q}}$. In line 17, the quantized reliable symbols are used to remove interference from the received signal vector via

$$\mathbf{y}^{(k+1)} = \mathbf{y}^{(k)} - \mathbf{G}_i(\sqrt{\rho_1}\tilde{\mathbf{x}}_{1,\text{q}} + \sqrt{\rho_2}\tilde{\mathbf{x}}_{2,\text{q}}). \qquad (15)$$

The quantized symbols are also stored in the estimated symbol vectors $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ (line 18). After canceling the interference from the detected symbols of both users, the algorithm updates the sets $\mathcal{N}_1$ and $\mathcal{N}_2$ of undetected symbols, and the set $\mathcal{D}_1$ of detected User 1 symbols, based on the state of the output vectors $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$. Since this is at the User i receiver, the algorithm stops when all of the User i symbols are detected, i.e., User 1 will detect all of its symbols before it detects all the User 2 symbols and can therefore stop once $\hat{\mathbf{x}}_1$ has no entries equal to zero.

Clearly, the performance of the RZ detector and the interference cancellation depend heavily on the thresholds T_1 and T_2. To the best of the authors' knowledge, in all existing works in the literature which use RZ detection, the thresholds are pre-determined and are reduced geometrically in each iteration [10], [19], [22]. However, in the NOMA context the performance of a user can be significantly affected by the MUI from the other user (especially for the user with lower power allocation). Hence, it is beneficial to optimize the thresholds T_1 and T_2 to improve the detection performance. Consequently, we use a greedy approach in which T_1 and T_2 are optimized within each iteration; for this, the post-equalization MSE, γ, is needed. The conventional LSQR algorithm of [23] does not provide this, and therefore a modified LSQR algorithm is proposed in the following section.

In this paper we focus on the 2-user case as this allows for greater simplicity and clarity in our analysis. However, while Algorithm 1 is presented for the case of 2 users, it can be modified in a straightforward manner to deal with the case of J users where $J \geq 2$, as follows. First, the sets $\mathcal{N}_j = \{n \in \mathcal{N} : \hat{x}_j[n] = 0\}$ and $\mathcal{D}_j = \mathcal{N} \backslash \mathcal{N}_j$ are defined for each User $j \in \mathcal{J}$, where $\mathcal{J} = \{1, 2, \ldots, J\}$. Second, Line 7 in Algorithm 2 is replaced by a loop which sets $\tilde{\mathbf{x}}_j = ((\tilde{\mathbf{x}}_{\text{sup}}[m])_{m \in \mathcal{N}_j \cap \mathcal{D}_{j-1}})/\sqrt{\rho_j}$ for each $j = 2$ to J. Finally, Lines 8-14 in Algorithm 1 are replaced by a loop where, for each $j \in \mathcal{J}$, threshold T_j is determined by solving (40) if $j = i$, by solving (44) if $j > i$, and by solving (48) if $j < i$. Here, the references to (44) and (48) refer to these optimization problems with appropriately modified user indices.

IV. Modified LSQR Algorithm

In this section, we present our proposed modified version of the LSQR algorithm, which is listed in Algorithm 2. We begin by summarizing the basic operation of the conventional LSQR algorithm, which remains unchanged in Algorithm 2. Then we describe the proposed modification which computes the post-equalization MSE. Two methods are presented for computing this MSE, an exact method and a low-complexity approximation.

A. Conventional LSQR algorithm

LSQR is a well-known iterative algorithm for solving equalization problems of the form $\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{w}$, where $\mathbf{x}$ is the transmitted vector, $\mathbf{y}$ is the received vector, $\mathbf{G}$ is the sparse channel matrix and $\mathbf{w}$ is the complex AWGN noise vector with variance per dimension σ^2 [24]. At iteration u, LSQR constructs the vector $\mathbf{x}_u$ in the Krylov subspace

$$\mathcal{K}(\mathbf{G}^{\text{H}}\mathbf{G}, \mathbf{G}^{\text{H}}\mathbf{y}, u) = \text{span}\{\mathbf{G}^{\text{H}}\mathbf{y}, (\mathbf{G}^{\text{H}}\mathbf{G})\mathbf{G}^{\text{H}}\mathbf{y}, \ \ldots \ , \\ (\mathbf{G}^{\text{H}}\mathbf{G})^{u-1}\mathbf{G}^{\text{H}}\mathbf{y}\}$$

which minimizes the norm of the residual, $||\mathbf{y} - \mathbf{G}\mathbf{x}_k||$. LSQR can also be regularized by including σ^2 as a damping parameter. After several iterations, LSQR provides performance similar to MMSE but with lower complexity [24]. At each iteration, the LSQR algorithm uses Golub-Kahan bidiagonalization and QR decomposition to obtain the estimate $\mathbf{x}_u$ [24]. The authors of [23] proposed a simple recursive method for updating this estimate within each iteration. The iterative process continues until either the norm of the residual reaches a pre-determined tolerance, ϵ, or the maximum number of iterations U is reached. The conventional implementation of LSQR does not compute the post-equalization MSE on the symbols in $\mathbf{x}_u$ which is necessary to optimize the thresholds of the RZ detector. In order to obtain the MSE, we propose to modify the LSQR algorithm to compute this directly within the LSQR process. In the following subsections, we present two methods for computing the MSE, an exact method and a novel low-complexity approximation.

B. Exact MSE computation

We note that LSQR is algebraically equivalent to applying the conjugate gradient (CG) method to the normal equation $\mathbf{G}^{\text{H}}\mathbf{G}\mathbf{x} = \mathbf{G}^{\text{H}}\mathbf{y}$ [25]. Therefore, we can adapt the method used in [26] for computing the post-equalization signal-to-interference-plus-noise ratio (SINR) of the CG method to LSQR.

LSQR computes $\mathbf{x}_u$ at each iteration using a simple recursion. However, similar to the CG method in [26], $\mathbf{x}_u$ can also

Algorithm 2 Modified LSQR Algorithm

```
1: Input: $\mathbf{G}$, $\mathbf{y}$ and $\sigma^2$
2: Initialize: $\mathbf{b} = \left(\begin{smallmatrix}\mathbf{y}\\ \mathbf{0}\end{smallmatrix}\right)$, $\mathbf{A} = \left(\begin{smallmatrix}\mathbf{G}\\ \sigma\mathbf{I}\end{smallmatrix}\right)$, $\beta_0 = \|\mathbf{b}\|$, $\mathbf{u}_0 = \mathbf{b}/\beta_0$, $\alpha_0 = \|\mathbf{A}^{\mathrm{H}}\mathbf{u}_0\|$, $\mathbf{v}_0 = \mathbf{A}^{\mathrm{H}}\mathbf{u}_0/\alpha_0$, $\mathbf{w}_0 = \mathbf{v}_0$, $\bar{\phi}_0 = \beta_0$, $\bar{\rho}_0 = \alpha_0$, $\mathbf{x}_0 = \mathbf{0}_{MN\times 1}$, $\mathbf{L}_1 = \frac{\tau_1}{\bar{\rho}_0\bar{\phi}_0}\mathbf{I}_{\mathrm{MN}}$, $\mathbf{L}_0 = \mathbf{0}_{MN\times MN}$, $\tau_0 = 1$ and $\bar{\phi}_u = \bar{\rho}_u = 1$ for $u < 0$
3: for $u = 1 : U$ do
4:     $\beta_u = \|\mathbf{A}\mathbf{v}_{u-1} - \alpha_{u-1}\mathbf{u}_{u-1}\|$
5:     $\mathbf{u}_u = (\mathbf{A}\mathbf{v}_{u-1} - \alpha_{u-1}\mathbf{u}_{u-1})/\beta_u$
6:     $\alpha_u = \|\mathbf{A}^{\mathrm{H}}\mathbf{u}_u - \beta_u\mathbf{v}_{u-1}\|$
7:     $\mathbf{v}_u = (\mathbf{A}^{\mathrm{H}}\mathbf{u}_u - \beta_u\mathbf{v}_{u-1})/\alpha_u$
8:     $\rho_u = \|[\bar{\rho}_{u-1} \;\; \beta_u]\|$, $c_u = \frac{\bar{\rho}_{u-1}}{\rho_u}$, $s_u = \frac{\beta_u}{\rho_u}$
9:     $\theta_u = s_u\alpha_u$, $\phi_u = c_u\bar{\phi}_{u-1}$
10:    $\tau_u = \frac{\phi_u}{\rho_u}$, $\mu_u = \frac{\theta_u}{\rho_u}$
11:    $\bar{\phi}_u = -s_u\bar{\phi}_{u-1}$, $\bar{\rho}_u = -c_u\alpha_u$
12:    $\mathbf{x}_u = \mathbf{x}_{u-1} + \tau_u\mathbf{w}_{u-1}$
13:    $\mathbf{w}_u = \mathbf{v}_u - \mu_u\mathbf{w}_{u-1}$
14:    Compute $\mathbf{L}_u$ using (21)
15:    if $||\mathbf{y} - \mathbf{G}\mathbf{x}_u|| \leq \epsilon$, break
16: end for
17: Compute $\psi[n]$ and $\nu[n]^2$, $\forall n$ using (22) and (23)
18: Compute $\gamma[n] = \frac{\nu[n]^2}{\psi[n]^2}$, $\forall n$
19: Output: $\tilde{\mathbf{x}} = \mathbf{x}_u$ and $\boldsymbol{\gamma}$
```

be computed using an LSQR equivalent equalization matrix which depends on the iteration index u. The LSQR equivalent equalization matrix at iteration u is defined as $\mathbf{L}_u\mathbf{G}^{\mathrm{H}}$, and $\mathbf{x}_u$ can be written as

$$\mathbf{x}_u = \mathbf{L}_u\mathbf{G}^{\mathrm{H}}\mathbf{y}. \tag{16}$$

If $\mathbf{L}_u$ is known, then the MSE on each symbol in $\mathbf{x}_u$ can be calculated. In the following, we derive a recursive method for computing $\mathbf{L}_u$ using variables which are already calculated within the LSQR process. From [26], note that the normal equation residual, $\boldsymbol{\xi}_u$, can be recursively calculated as

$$\boldsymbol{\xi}_u = \boldsymbol{\xi}_{u-1} - \tau_u\mathbf{A}^{\mathrm{H}}\mathbf{A}\mathbf{w}_{u-1}, \tag{17}$$

where $\mathbf{A} = \left(\begin{smallmatrix}\mathbf{G}\\ \sigma\mathbf{I}\end{smallmatrix}\right)$. This can also be calculated as [26]

$$\boldsymbol{\xi}_u = \bar{\phi}_u\bar{\rho}_u\mathbf{w}_u - \mu_u^2\bar{\phi}_{u-1}\bar{\rho}_{u-1}\mathbf{w}_{u-1}. \tag{18}$$

We then substitute $\boldsymbol{\xi}_u$ from (18) into (17) to obtain

$$\begin{aligned}\bar{\phi}_u\bar{\rho}_u\mathbf{w}_u =& \mu_u^2\bar{\phi}_{u-1}\bar{\rho}_{u-1}\mathbf{w}_{u-1} + \bar{\phi}_{u-1}\bar{\rho}_{u-1}\mathbf{w}_{u-1}\\ &- \mu_{u-1}^2\bar{\phi}_{u-2}\bar{\rho}_{u-2}\mathbf{w}_{u-2} - \tau_u\mathbf{A}^{\mathrm{H}}\mathbf{A}\mathbf{w}_{u-1}\end{aligned} \tag{19}$$

Next, we rewrite line 12 of Algorithm 2 as $\mathbf{w}_{u-1} = (\mathbf{x}_u - \mathbf{x}_{u-1})/\tau_u$ which can then be substituted into (19) to obtain the following recursion for $\mathbf{x}_u$:

$$\begin{aligned}\mathbf{x}_u =& \mathbf{x}_{u-1} +\\ &\left(\frac{\tau_u\bar{\rho}_{u-2}\bar{\phi}_{u-2}(1+\mu_{u-1}^2)}{\tau_{u-1}\bar{\rho}_{u-1}\bar{\phi}_{u-1}}\mathbf{I}_{\mathrm{MN}} - \frac{\tau_u}{\bar{\rho}_{u-1}\bar{\phi}_{u-1}}\mathbf{A}^{\mathrm{H}}\mathbf{A}\right)\\ &\times(\mathbf{x}_{u-1} - \mathbf{x}_{u-2})\\ &+ \frac{\mu_{u-2}^2\tau_u\bar{\rho}_{u-3}\bar{\phi}_{u-3}}{\tau_{u-2}\bar{\rho}_{u-1}\bar{\phi}_{u-1}}(\mathbf{x}_{u-2} - \mathbf{x}_{u-3}).\end{aligned} \tag{20}$$

Using (16), we can obtain the recursion for $\mathbf{L}_u$ as

$$\begin{aligned}\mathbf{L}_u =& \mathbf{L}_{u-1} +\\ &\left(\frac{\tau_u\bar{\rho}_{u-2}\bar{\phi}_{u-2}(1+\mu_{u-1}^2)}{\tau_{u-1}\bar{\rho}_{u-1}\bar{\phi}_{u-1}}\mathbf{I}_{\mathrm{MN}} - \frac{\tau_u}{\bar{\rho}_{u-1}\bar{\phi}_{u-1}}\mathbf{A}^{\mathrm{H}}\mathbf{A}\right)\\ &\times(\mathbf{L}_{u-1} - \mathbf{L}_{u-2})\\ &+ \frac{\mu_{u-2}^2\tau_u\bar{\rho}_{u-3}\bar{\phi}_{u-3}}{\tau_{u-2}\bar{\rho}_{u-1}\bar{\phi}_{u-1}}(\mathbf{L}_{u-2} - \mathbf{L}_{u-3}),\end{aligned} \tag{21}$$

where we initialize $\mathbf{L}_1 = \frac{\tau_1}{\bar{\rho}_0\bar{\phi}_0}\mathbf{I}_{\mathrm{MN}}$, $\mathbf{L}_u = \mathbf{0}_{MN\times MN}$ for $u \leq 0$, $\tau_0 = 1$ and $\bar{\phi}_u = \bar{\rho}_u = 1$ for $u < 0$.

The matrix $\mathbf{L}_u$ can now be used to compute the MSE. Let $\mathbf{B} = \mathbf{L}_u\mathbf{Z}$, where $\mathbf{Z} = \mathbf{G}^{\mathrm{H}}\mathbf{G}$. The post-equalization channel gain on element n of $\mathbf{x}_u$ is given by

$$\psi[n] = B[n,n]. \tag{22}$$

The variance of the interference-plus-noise on element n of $\mathbf{x}_u$ is given by

$$\nu[n]^2 = \sum_{m,m\neq n}|B[n,m]|^2 + C[n,n]\sigma^2, \tag{23}$$

where $\mathbf{C} = \mathbf{B}\mathbf{L}_u^{\mathrm{H}}$. The MSE of element n of $\mathbf{x}_u$ is therefore given by

$$\gamma[n] = \frac{\nu[n]^2}{\psi[n]^2}. \tag{24}$$

While this method provides the exact MSE of each symbol at iteration u of the LSQR process, it is computationally complex due to the $MN \times MN$ matrix multiplication in (21) which requires $(MN)^2$ complex multiplications. In the next subsection, we propose a approximation to this MSE which has a significantly lower computational complexity.

C. Low-complexity approximation

In practice, it is impossible to estimate the channel gains at each individual time sample n, i.e., all of the values of $h[n,l]$. Thus, we assume that the channel is varying sufficiently slowly that it has an approximately constant CIR over each OFDM symbol within an OTFS block. Under this condition, $\mathbf{G}$ is approximately a block circulant matrix with circulant blocks (BCCB) [21]. Therefore, $\mathbf{G}$ can be diagonalized via

$$\boldsymbol{\Lambda}_{\mathbf{G}} = (\mathbf{F}_N \otimes \mathbf{F}_M)\mathbf{G}(\mathbf{F}_N \otimes \mathbf{F}_M)^{\mathrm{H}}. \tag{25}$$

The matrix $\mathbf{A}^{\mathrm{H}}\mathbf{A}$ inherits the BCCB structure of $\mathbf{G}$. Therefore, we can obtain the diagonalization of $\mathbf{A}^{\mathrm{H}}\mathbf{A}$ as

$$\boldsymbol{\Lambda}_{\mathbf{A}} = (\mathbf{F}_N \otimes \mathbf{F}_M)\mathbf{A}^{\mathrm{H}}\mathbf{A}(\mathbf{F}_N \otimes \mathbf{F}_M)^{\mathrm{H}}.$$

By using the properties of BCCB matrices [5], we can alternatively obtain $\boldsymbol{\Lambda}_{\mathbf{A}}$ as

$$\boldsymbol{\Lambda}_{\mathbf{A}} = (\boldsymbol{\Lambda}_{\mathbf{G}}{}^*\boldsymbol{\Lambda}_{\mathbf{G}} + \sigma^2\mathbf{I}).$$

Note that $\mathbf{L}_1$ is initialized as a diagonal matrix and hence $\mathbf{L}_u$ retains the BCCB structure of the $\mathbf{A}^{\mathrm{H}}\mathbf{A}$ for $u > 2$. Additionally, since $\mathbf{L}_1$ is a diagonal matrix, it is invariant under diagonalization, i.e,

$$\boldsymbol{\Lambda}_{\mathbf{L}_1} = (\mathbf{F}_N \otimes \mathbf{F}_M)\mathbf{L}_1(\mathbf{F}_N \otimes \mathbf{F}_M)^{\mathrm{H}} = \mathbf{L}_1.$$

This means that the entire recursion can be performed in the diagonalized domain. The recursion for $\mathbf{\Lambda}_{\mathbf{L}_u}$ can now be formulated as

$$\begin{aligned}\mathbf{\Lambda}_{\mathbf{L}_u} =& \mathbf{\Lambda}_{\mathbf{L}_{u-1}} + \\ & \left(\frac{\tau_u \bar{\rho}_{u-2} \bar{\phi}_{u-2}(1+\mu_{u-1}^2)}{\tau_{u-1}\bar{\rho}_{u-1}\bar{\phi}_{u-1}} \mathbf{I}_{\mathrm{MN}} - \frac{\tau_u}{\bar{\rho}_{u-1}\bar{\phi}_{u-1}} \mathbf{\Lambda}_{\mathbf{A}}\right) \\ & \times \left(\mathbf{\Lambda}_{\mathbf{L}_{u-1}} - \mathbf{\Lambda}_{\mathbf{L}_{u-2}}\right) \\ & + \frac{\mu_{u-2}^2 \tau_u \bar{\rho}_{u-3} \bar{\phi}_{u-3}}{\tau_{u-2}\bar{\rho}_{u-1}\bar{\phi}_{u-1}} \left(\mathbf{\Lambda}_{\mathbf{L}_{u-2}} - \mathbf{\Lambda}_{\mathbf{L}_{u-3}}\right). \end{aligned} \tag{26}$$

where we initialize $\mathbf{\Lambda}_{\mathbf{L}_1} = \frac{\tau_1}{\bar{\rho}_0 \bar{\phi}_0} \mathbf{I}_{\mathrm{MN}}$ and $\mathbf{\Lambda}_{\mathbf{L}_u} = \mathbf{0}_{MN \times MN}$ for $u < 1$. Since this recursion only involves diagonal matrices, it can be performed with low complexity.

We can now use $\mathbf{\Lambda}_{\mathbf{L}_u}$ to calculate the approximate MSE. We calculate the diagonalizations of $\mathbf{B}$ and $\mathbf{C}$ as $\mathbf{\Lambda}_{\mathbf{B}} = \tilde{\mathbf{L}} \mathbf{\Lambda}_{\mathbf{G}}{}^* \mathbf{\Lambda}_{\mathbf{G}}$ and $\mathbf{\Lambda}_{\mathbf{C}} = \mathbf{\Lambda}_{\mathbf{G}}{}^* \mathbf{\Lambda}_{\mathbf{G}} \mathbf{\Lambda}_{\mathbf{L}_u}$, respectively. The reverse of the diagonalization process in (25) can then be used to calculate approximations of $\mathbf{B}$ and $\mathbf{C}$ as

$$\tilde{\mathbf{B}} = (\mathbf{F}_N \otimes \mathbf{F}_M)^{\mathrm{H}} \mathbf{\Lambda}_{\mathbf{B}} (\mathbf{F}_N \otimes \mathbf{F}_M)$$

and

$$\tilde{\mathbf{C}} = (\mathbf{F}_N \otimes \mathbf{F}_M)^{\mathrm{H}} \mathbf{\Lambda}_{\mathbf{C}} (\mathbf{F}_N \otimes \mathbf{F}_M).$$

Since $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{C}}$ are BCCB matrices, their respective rows are simply shifted versions of each other. Therefore, under this approximation, each symbol experiences the same MSE and the subscript n can be dropped from (22) – (24). The post-equalization channel gain is simply given by

$$\tilde{\psi} = \tilde{B}[1,1], \tag{27}$$

and the variance of the interference-plus-noise is given by

$$\tilde{\nu}^2 = \sum_{m=2}^{MN-1} |\tilde{B}[1,m]|^2 + \tilde{C}[1,1]\sigma^2. \tag{28}$$

Therefore, the approximate MSE on each symbol is obtained as

$$\tilde{\gamma} = \frac{\tilde{\nu}^2}{\tilde{\psi}^2}. \tag{29}$$

In the context of the considered OTFS-NOMA system, we apply mLSQR in line 5 of Algorithm 1 to obtain in the k-th iteration the estimate of the transmitted superimposed symbol vector, $\tilde{\mathbf{x}}_{\mathrm{sup}}$ and the post-equalization MSE for User 1 and User 2, which is given by

$$\tilde{\gamma}_j^{(k)} = \frac{\tilde{\nu}^2}{\rho_j \tilde{\psi}^2}, \forall j \in \{1,2\} \tag{30}$$

We then use this calculated MSE to optimize the thresholds of the RZ detector in a greedy manner, as described in detail in the following section.

It is important to note that the proposed modifications to the LSQR algorithm do not change the computational procedure of LSQR; instead, the modifications use terms that are already calculated in LSQR to obtain the post-equalization MSE. As such, the proposed modifications do not affect the numerical stability of LSQR.

V. RZ Detector Threshold Optimization

In this section, we describe how the MSE calculated by the modified LSQR algorithm can be used to optimize the RZ thresholds for each user. Our proposed method works by tracking the evolution of the MSE on the symbols of User 1 and User 2 as Algorithm 1 progresses. The key idea is to choose optimal values for the RZ thresholds T_1 and T_2 in each iteration k which minimise the "pre-decision" MSE, i.e., the mean-square value of the error seen by the RZ detector at iteration $k+1$.

In lines 15 and 16 of Algorithm 1, at iteration k, the RZ detector makes a decision on whether the received symbols of each user are unreliable or reliable and then quantizes the reliable symbols to the nearest QAM symbol in that user's constellation. Therefore, there are 3 possible outcomes of the unreliable zone detection. Symbols are either correct, incorrect or undetected, each such event having its own associated probability which depends on the thresholds, T_1 and T_2, and the user's MSE values, $\tilde{\gamma}_1^k$ and $\tilde{\gamma}_2^k$.

Next, we derive expressions for the probability of each outcome above in the context of each user's symbols. To derive the probability of each outcome for a generalized A_i-ary QAM system, we first derive them for a $\sqrt{A_i}$-ary PAM system by adapting the closed-form expression for the probability of error of a 2-user NOMA system derived in [27]. For the User 1 symbols, the decision is being made on the superimposed symbols which contain contributions from the symbols of User 1 *and* User 2. First, the following functions are defined (c.f. [27]):

$$\mathrm{q_a}(j,l,T_1,\tilde{\gamma}_1) = \mathrm{Q}\left(\frac{d((2j-1)-(2l-1)\sqrt{\frac{\rho_2}{\rho_1}}) - \frac{T_1}{2}}{\sqrt{\tilde{\gamma}_1/2}}\right),$$

$$\mathrm{q_b}(j,l,T_1,\tilde{\gamma}_1) = \mathrm{Q}\left(\frac{d((2j-1)+(2l-1)\sqrt{\frac{\rho_2}{\rho_1}}) - \frac{T_1}{2}}{\sqrt{\tilde{\gamma}_1/2}}\right),$$

$$\mathrm{q_c}(l,T_1,\tilde{\gamma}_1) = \mathrm{Q}\left(\frac{d(1-(2l-1)\sqrt{\frac{\rho_2}{\rho_1}}) + \frac{T_1}{2}}{\sqrt{\tilde{\gamma}_1/2}}\right),$$

$$\mathrm{q_d}(l,T_1,\tilde{\gamma}_1) = \mathrm{Q}\left(\frac{d(1+(2l-1)\sqrt{\frac{\rho_2}{\rho_1}}) + \frac{T_1}{2}}{\sqrt{\tilde{\gamma}_1/2}}\right),$$

where $\mathrm{Q}(x) = \frac{1}{2}\mathrm{erfc}(\frac{x}{\sqrt{2}})$ denotes the Gaussian Q-function. The probability of correct symbol detection per dimension for User 1, denoted by $P_{\mathrm{C,PAM,1}}$, is then given by (31), shown at the top of the next page, where the threshold used is T_1. For the probability of incorrect detection, we adopt a *nearest-neighbor approximation*, i.e., it is assumed that if an incorrect symbol is detected, it is always a nearest neighbor in that user's QAM constellation (this assumption becomes very accurate at high SNR). The probability of incorrect detection per dimension for User 1, denoted by $P_{\mathrm{E,PAM,1}}$, is given by (32), where the threshold used is T_1. User 2 symbols are only fed into the RZ detector once the corresponding User 1 symbols have been detected on a previous iteration. Thereafter, the decisions

$$P_{\mathrm{C,PAM},1} = 1 - \frac{2(\sqrt{A_1}-1)}{A_1}\sum_{l=1}^{\sqrt{A_1}/2}\left[\mathrm{q_a}(1,l,T_1,\tilde{\gamma}_1)+\mathrm{q_b}(1,l,T_1,\tilde{\gamma}_1)\right], \tag{31}$$

$$P_{\mathrm{E,PAM},1} = \frac{2}{A}\left(\sum_{l=1}^{\sqrt{A_1}/2}\Big[(\sqrt{A_1}-1)\left(\mathrm{q_c}(1,l,T_1,\tilde{\gamma}_1)+\mathrm{q_d}(1,l,T_1,\tilde{\gamma}_1)\right) - (\sqrt{A_1}-2)\left(\mathrm{q_a}(2,l,T_1,\tilde{\gamma}_1)+\mathrm{q_b}(2,l,T_1,\tilde{\gamma}_1)\right)\Big]\right) \tag{32}$$

$$P_{\mathrm{C,PAM},2} = 1 - \frac{2(\sqrt{A_2}-1)}{\sqrt{A_2}}\mathrm{Q}\left(\frac{d-\frac{T_2}{2}}{\sqrt{\tilde{\gamma}_2/2}}\right), \tag{33}$$

$$P_{\mathrm{E,PAM},2} = \frac{2}{\sqrt{A_2}}\left((\sqrt{A_2}-1)\mathrm{Q}\left(\frac{d+\frac{T_2}{2}}{\sqrt{\tilde{\gamma}_2/2}}\right) - (\sqrt{A_2}-2)\mathrm{Q}\left(\frac{3d-\frac{T_2}{2}}{\sqrt{\tilde{\gamma}_2/2}}\right)\right) \tag{34}$$

are no longer being made upon a superposition of both user symbols. Consequently, the probability of correct detection per dimension of User 2 is given by (33), where the threshold used is T_2. The probability of incorrect detection per dimension of User 2 is given by (34). The probability of correct detection, incorrect detection, and non-detection at User $i \in \{1, 2\}$ can then be obtained, respectively, as

$$P_{\mathrm{c},i} = (P_{\mathrm{C,PAM},i})^2, \tag{35}$$

$$P_{\mathrm{e},i} = 2P_{\mathrm{E,PAM},i}, \tag{36}$$

$$P_{\mathrm{u},i} = 1 - P_{\mathrm{c},i} - P_{\mathrm{e},i}, \tag{37}$$

The next subsection describes how these probability expressions can be used to optimize the thresholds at each user's receiver.

A. Design of threshold T_i at User i receiver

In this subsection we describe the process for choosing the threshold T_i at the receiver of User i in each iteration of Algorithm 1. We begin at iteration $k = 1$ where it can be seen from (27), (28) and (29) the MSE of the User i symbols after mLSQR equalization is given by

$$\tilde{\gamma}_i^{(1)} = \frac{1}{\rho_i\tilde{\psi}^2}\left(\sum_{m=2}^{MN-1}|\tilde{B}[1,m]|^2(\rho_1+\rho_2) + \tilde{C}[1,1]\sigma^2\right). \tag{38}$$

This can be rewritten as

$$\tilde{\gamma}_i^{(1)} = \Omega_i^{(1)} + \Psi_{i,u}^{(1)} + W_i. \tag{39}$$

where $\Omega_i^{(1)} = \frac{\rho_1}{\rho_i\tilde{\psi}^2}\sum_{m=2}^{MN-1}|\tilde{B}[1,m]|^2$ is the MSE due to the undetected User 1 symbols, $\Psi_{i,u}^{(1)} = \frac{\rho_2}{\rho_i\tilde{\psi}^2}\sum_{m=2}^{MN-1}|\tilde{B}[1,m]|^2$ is the MSE due to the undetected User 2 symbols and W_i is the AWGN component of the MSE. After the detection and interference cancellation process in lines 15–17 of Algorithm 1, the MSE due to undetected User 1 symbols will be reduced by a factor depending on the probability of non-detection of User 1 symbols in iteration 1. Therefore, we can express the remaining MSE of the undetected User 1 symbols at iteration $k = 2$ as $\Omega_i^{(2)} = \Omega_i^{(1)}P_{\mathrm{u},1}^{(1)}$. Generalizing this argument, at iteration k, we express the remaining MSE of the undetected User 1 symbols as $\Omega_i^{(k)} = \Omega_i^{(k-1)}P_{\mathrm{u},1}^{(k-1)}$ and we define the remaining MSE of the undetected User 2 symbols as $\Psi_{i,u}^{(k)} = \Psi_{i,u}^{(k-1)}P_{\mathrm{u},2}^{(k-2)}$. Since W_i is unaffected by the interference cancellation process, it sets a limit on the minimum achievable probability of error. Therefore, the User i receiver should choose a threshold at iteration k which achieves this minimum minimum achievable probability of error. We note that, via (36), (32) and (34), $P_{\mathrm{e},i}$ can be expressed as a function of 2 variables, i.e, $\tilde{\gamma}_i$ and T_i. Hence, the User i receiver chooses the threshold T_i at iteration k such that

$$P_{\mathrm{e},i}(\tilde{\gamma}_i^{(k)}, T_i) = P_{\mathrm{e},i}(W_i, 0). \tag{40}$$

Since the remaining User j symbols ($j \neq i$) impart MUI on the remaining undetected User i symbols, User i must select the threshold T_j which minimizes $\tilde{\gamma}_i^{(k)}$. We now describe the exact optimization process at each user's receiver.

B. Optimizing T_2 at receiver of User 1

The MSE of User 1 will be reduced by the correctly detected symbols from the previous iteration and increased by the incorrectly detected symbols. Hence, the MSE of User 1 at iteration 2 will be comprised of the remaining interference from the undetected User 1 symbols, the interference from the undetected User 2 symbols, the AWGN and the MSE due to interference cancellation error multiplied by the probability of error of User 1. Therefore, the MSE for User 1 at iteration 2 is given by

$$\tilde{\gamma}_1^{(2)} = \Omega_1^{(1)}P_{\mathrm{u},1}^{(1)} + E_1\Omega_1^{(1)}P_{\mathrm{e},1}^{(1)} + \Psi_{1,u}^{(1)} + W, \tag{41}$$

where $E_i = 4d_i^2$ is the average energy of an interfering symbol due to the event of interference cancellation error of User i under the nearest-neighbour approximation. This is multiplied by the probability of error of User i and by the remaining MSE due to undetected User i symbols to account for the reduced number of symbols can be incorrectly detected as Algorithm 1 progresses. Using (37), we can express (41) as

$$\tilde{\gamma}_1^{(2)} = \tilde{\gamma}_1^{(1)} - \Omega_1^{(1)}P_{\mathrm{c},1}^{(1)} + (E_1-1)\Omega_1^{(1)}P_{\mathrm{e},1}^{(1)}, \tag{42}$$

$$\begin{aligned}\frac{\partial}{\partial T_1}P_{\mathrm{c},1}^{(k)} = &\left(2-\frac{4(\sqrt{A_1}-1)}{A_1}\sum_{l=1}^{\sqrt{A_1}/2}[\mathrm{q_a}(1,l)+\mathrm{q_b}(1,l)]\right)\times \\ &\left(\frac{(\sqrt{A_1}-1)}{A_1\sqrt{\pi\tilde{\gamma}_1^{(k)}}}\sum_{l=1}^{\sqrt{A_1}/2}\left[\exp\left(-\frac{(d((4l-2)\sqrt{\frac{\rho_2}{\rho_1}}-2)+T_1)^2}{4\tilde{\gamma}_1^{(k)}}\right)+\exp\left(-\frac{(d((2-4l)\sqrt{\frac{\rho_2}{\rho_1}}-2)+T_1)^2}{4\tilde{\gamma}_1^{(k)}}\right)\right]\right)\end{aligned} \tag{50}$$

$$\begin{aligned}\frac{\partial}{\partial T_1}P_{\mathrm{e},1}^{(k)} = &\frac{-2}{A_1\sqrt{\pi\tilde{\gamma}_1^{(k)}}}\times \\ &\left(\sum_{l=1}^{\sqrt{A_1}/2}\left[(\sqrt{A_1}-1)\left(\exp\left(-\frac{(d((2-4l)\sqrt{\frac{\rho_2}{\rho_1}}+2)+T_1)^2}{4\tilde{\gamma}_1^{(k)}}\right)+\exp\left(-\frac{(d((4l-2)\sqrt{\frac{\rho_2}{\rho_1}}+2)+T_1)^2}{4\tilde{\gamma}_1^{(k)}}\right)\right)\right.\right. \\ &\left.\left.+(\sqrt{A_1}-2)\left(\exp\left(-\frac{(d((4l-2)\sqrt{\frac{\rho_2}{\rho_1}}-6)+T_1)^2}{4\tilde{\gamma}_1^{(k)}}\right)+\exp\left(-\frac{(d((2-4l)\sqrt{\frac{\rho_2}{\rho_1}}-6)+T_1)^2}{4\tilde{\gamma}_1^{(k)}}\right)\right)\right]\right)\end{aligned} \tag{51}$$

$$\frac{\partial}{\partial T_2}P_{\mathrm{c},2}^{(k)} = \left(2-\frac{4(\sqrt{A_2}-1)}{A_2}\mathrm{Q}\left(\frac{d-\frac{T_2}{2}}{\sqrt{\tilde{\gamma}_2^{(k)}/2}}\right)\right)\left(\frac{(\sqrt{A_2}-1)}{A_2\sqrt{\pi\tilde{\gamma}_2^{(k)}}}\exp\left(-\frac{(2d+T_2)^2}{4\tilde{\gamma}_2^{(k)}}\right)\right) \tag{52}$$

$$\frac{\partial}{\partial T_2}P_{\mathrm{e},2}^{(k)} = \frac{-1}{A_2\sqrt{\pi\tilde{\gamma}_1^{(k)}}}\left((\sqrt{A_2}-1)\exp\left(-\frac{(2d+T_2)^2}{4\tilde{\gamma}_2^{(k)}}\right)+(\sqrt{A_2}-2)\exp\left(-\frac{(T_2-6d)^2}{4\tilde{\gamma}_2^{(k)}}\right)\right) \tag{53}$$

Generalizing this argument, we can formulate an expression for the MSE of User 1 on iteration $k+1$, which is given by

$$\begin{aligned}\tilde{\gamma}_1^{(k+1)} &= \tilde{\gamma}_1^{(k)} - \Omega_1^{(k)}P_{\mathrm{c},1}^{(k)} + (E_1-1)\Omega_1^{(k)}P_{\mathrm{e},1}^{(k)} \\ &- \left(\Psi_{1,u}^{(k)}\left(P_{\mathrm{c},1}^{(k-1)}+P_{\mathrm{e},1}^{(k-1)}\right)+\Psi_{1,d}^{(k)}\right)P_{\mathrm{c},2}^{(k)} \\ &+ (E_2-1)\left(\Psi_{1,u}^{(k)}\left(P_{\mathrm{c},1}^{(k-1)}+P_{\mathrm{e},1}^{(k-1)}\right)+\Psi_{1,d}^{(k)}\right)P_{\mathrm{e},2}^{(k)}.\end{aligned} \tag{43}$$

where $\Psi_{1,d}^{(k)} = \Psi_{1,d}^{(k-1)} + \Psi_{1,u}^{(k-1)}\left(P_{\mathrm{c},1}^{(k-2)}+P_{\mathrm{e},1}^{(k-2)}\right)P_{\mathrm{u},2}^{(k-1)} - \Psi_{1,d}^{(k-1)}\left(P_{\mathrm{c},2}^{(k-1)}+P_{\mathrm{e},2}^{(k-1)}\right)$ is the remaining interference power from the User 2 symbols for which the corresponding User 1 symbols have been detected.

The probability of User 1 symbols being undetected is initialized as $P_{\mathrm{u},1}^{(0)} = 1$. We also initialize $\Psi_{1,d}^{(-1)} = 0$ as no User 1 symbols have been detected before the algorithm begins. The MSE for User 1 on iteration $k+1$ is a function of the probability terms in (35) and (36), which are themselves functions of T_2. All other terms are constants which can be updated recursively. The User 1 receiver can now choose the optimum value T_2 at iteration k to minimize the MSE of User 1 at iteration $k+1$.

At each iteration k, the receiver of User 1 solves the optimization problem

$$\min_{T_2} \quad \tilde{\gamma}_1^{(k+1)} \tag{44a}$$

$$\text{s.t.} \quad T_2 \geq 0. \tag{44b}$$

To solve this optimization problem, the derivative of $\tilde{\gamma}_1^{(k+1)}$ with respect to T_2 is set equal to zero. Since only $P_{\mathrm{e},2}^{(k)}$ and $P_{\mathrm{c},2}^{(k)}$ in (43) are functions of T_2, the derivative of $\tilde{\gamma}_1^{(k+1)}$ with respect to T_2 is given by

$$\begin{aligned}\frac{\partial}{\partial T_2}\tilde{\gamma}_1^{(k+1)} &= \left(\Psi_{2,u}^{(k)}P_{\mathrm{d},1}^{(k-1)}+\Psi_{1,d}^{(k)}\right)\times \\ &\left((E_2-1)\frac{\partial}{\partial T_2}P_{\mathrm{e},2}^{(k)} - \frac{\partial}{\partial T_2}P_{\mathrm{c},2}^{(k)}\right).\end{aligned} \tag{45}$$

Using $\frac{\partial}{\partial x}\mathrm{Q}(x) = -\frac{1}{\sqrt{2\pi}}e^{-x^2}$, the derivative of $P_{\mathrm{c},2}^{(k)}$ and $P_{\mathrm{e},2}^{(k)}$ can be expressed as (52) and (53), respectively. The User 1 receiver then solves $\frac{\partial}{\partial T_2}\tilde{\gamma}_1^{(k+1)} = 0$ using the Brent-Dekker method [28] to obtain the solution to (44a), which is the optimized T_2 at iteration k of Algorithm 1.

C. Optimizing T_1 at User 2 receiver

As with User 1, the MSE of User 2 will be reduced by the correctly detected symbols from the previous iteration and increased by the incorrectly detected symbols. However, in contrast to User 1, the RZ detector of User 2 only makes decisions on the User 2 symbols whose corresponding User 1 symbols have already been detected. Therefore, the MSE of a User 2 symbol is also affected by the incorrect detection of the overlapping User 1 symbol. Given this, at iteration 2 of

Algorithm 1, the MSE of the User 2 symbols which are being fed into the RZ detector can be written as

$$
\begin{aligned}
\tilde{\gamma}_2^{(2)} =& \tilde{\gamma}_2^{(1)} - \Omega_2^{(1)} P_{\mathrm{c},1}^{(1)} + (E_1 - 1)\Omega_2^{(1)} P_{\mathrm{e},1}^{(1)} \\
&+ \frac{\rho_2}{\rho_1} E_1 P_{\mathrm{u},1}^{(0)} P_{\mathrm{e},1}^{(1)} + W,
\end{aligned}
\tag{46}
$$

where the fourth term of (46) accounts for the MSE due to directly overlapping User 1 symbols that are incorrectly detected. Generalizing this argument, we can formulate a general expression for the MSE of User 2 at iteration $k+1$, which is given by

$$
\begin{aligned}
\tilde{\gamma}_2^{(k+1)} =& \tilde{\gamma}_2^{(k)} - \Omega_2^{(k)} P_{\mathrm{c},1}^{(k)} \\
&+ ((E_1 - 1)\Omega_2^{(k)} + \frac{\rho_2}{\rho_1} E_1 P_{\mathrm{u},1}^{(1)}) P_{\mathrm{e},1}^{(k)} \\
&- \left(\Psi_{2,u}^{(k)} P_{\mathrm{d},1}^{(k-1)} + \Psi_{2,d}^{(k)}\right) P_{\mathrm{c},2}^{(k)} \\
&+ (E_2 - 1)\left(\Psi_{2,u}^{(k)} P_{\mathrm{d},1}^{(k-1)} + \Psi_{2,d}^{(k)}\right) P_{\mathrm{e},2}^{(k)}.
\end{aligned}
\tag{47}
$$

Similar to the case of the User 1 receiver above, we can now formulate the optimization problem to be solved at iteration k, i.e.,

$$
\min_{T_1} \quad \tilde{\gamma}_2^{(k+1)} \tag{48a}
$$

$$
\text{s.t.} \quad T_1 \geq 0, \tag{48b}
$$

This optimization problem is solved in a similar manner to (44a). Since only $P_{\mathrm{e},1}^{(k)}$ and $P_{\mathrm{c},1}^{(k)}$ in (29) are functions of T_1, the derivative of $\tilde{\gamma}_2^{(k+1)}$ with respect to T_1 is given by

$$
\begin{aligned}
\frac{\partial}{\partial T_1} \tilde{\gamma}_1^{(k+1)} =& \left((E_1 - 1)\Omega_2^{(k)} + \frac{\rho_2}{\rho_1} E P_{\mathrm{u},1}^{(1)}\right) \frac{\partial}{\partial T_1} P_{\mathrm{e},1}^{(k)} \\
&- \Omega_2^{(k)} \frac{\partial}{\partial T_1} P_{\mathrm{c},1}^{(k)}
\end{aligned}
\tag{49}
$$

The derivatives of $P_{\mathrm{c},2}^{(k)}$ and $P_{\mathrm{e},2}^{(k)}$ can be expressed as (50) and (51) respectively. The receiver of User 2 then solves $\frac{\partial}{\partial T_2} \tilde{\gamma}_1^{(k+1)} = 0$ to obtain the solution to (48a), which is the optimized T_1 at iteration k of Algorithm 1.

Computational complexity

In this subsection, the computational complexity of the proposed method is compared to that of the MMSE-SIC benchmark, in terms of the number of complex multiplications. Direct implementation of MMSE equalization involves the inversion of an $MN \times MN$ matrix and hence has a computational complexity of $\mathcal{O}(M^3N^3)$. Each iteration of the conventional LSQR algorithm has a computational complexity of $\mathcal{O}(MN\log_2(MN))$ [24]. The low-complexity MSE calculation in the proposed mLSQR algorithm (described in Subsection IV-C) can be performed with a single M-point FFT operation and a single N-point IFFT operation and therefore has a computational complexity of $\mathcal{O}(M\log_2(M)) + \mathcal{O}(N\log_2(N))$, which is negligible compared to the complexity of the LSQR computation. Hence, the proposed mLSQR algorithm has a computational complexity of $\mathcal{O}(MN\log_2(MN))$. In the worst-case scenario, Algorithm 1 performs mLSQR K times, each with U mLSQR iterations; therefore, the computational complexity of Algorithm 1 is $\mathcal{O}(UKMN\log_2(MN))$. In practice, the typical values of K and U are in the order of tens and the typical values of M and N can be as high as $M = 512$ and $N = 128$ [4], [7]. Thus, $UK \ll M^2N^2$ and our method can achieve orders of magnitude computational complexity improvement over MMSE-SIC for OTFS-NOMA. It should also be noted that optimizing the thresholds allows for Algorithm 1 to converge faster at high SNR than a naive threshold design, as the thresholds are not unnecessarily large and the algorithm can detect more symbols at earlier iterations.

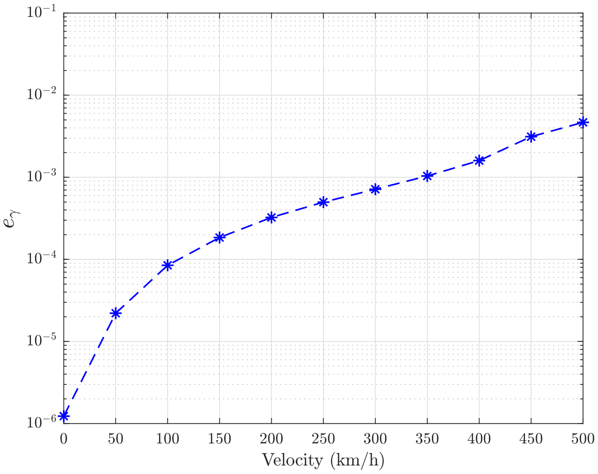

Fig. 2. Normalized approximation error of the proposed low-complexity MSE computation.

VI. Numerical Results and Discussion

TABLE I
Simulation Parameters

Delay bins (M)	64
Doppler bins (N)	16
Carrier frequency (f_c)	5.9 GHz
Subcarrier spacing	15 kHz
Modulation scheme	4-QAM, 16-QAM
Channel model	TDL-C [29]
Delay spread	300 ns
User velocity	$90 - 450$ km/h
Algorithm 1 iterations (K)	10
mLSQR iterations (U)	15
mLSQR tolerance (ϵ)	10^{-2}

This section presents numerical results to showcase the effectiveness of the proposed OTFS-NOMA equalization and detection algorithm. As a benchmark, an OTFS-NOMA system using MMSE equalization and SIC for detection is considered, which is referred to as MMSE-SIC. Additionally, the performance of the proposed algorithm using the optimized thresholds outlined in Section V is compared to the proposed algorithm with naive (conventional) threshold design. For the naive threshold case, we consider a starting threshold of $T_i^{(1)} = 2d_i$ for each user which is then reduced geometrically within each iteration as $T_i^{(k)} = T_i^{(1)}(1 - (k/K))$ (this was the threshold adaptation strategy adopted in [10], [19], [22]).

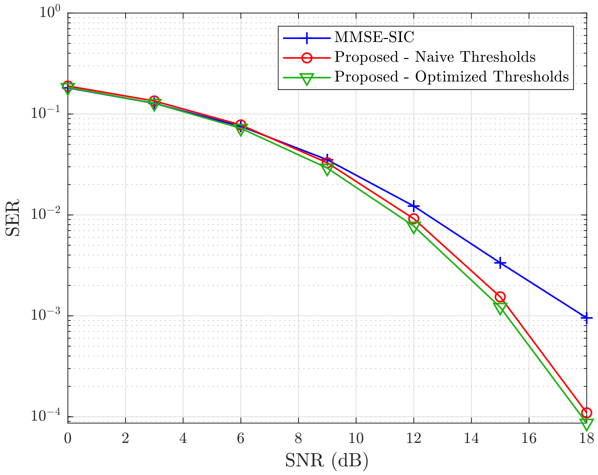

Fig. 3. Comparison of the SER performance of User 1 using Algorithm 1 with optimized thresholds, Algorithm 1 with naive thresholds, and MMSE equalization with SIC, with different SNR levels, for the case where each user is allocated a 4-QAM constellation.

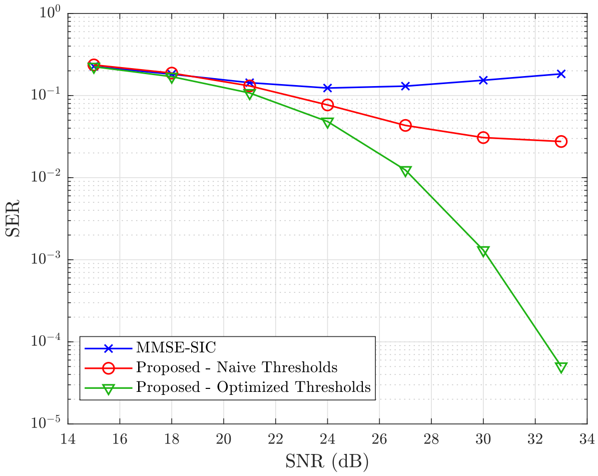

Fig. 4. Comparison of the SER performance of User 2 using Algorithm 1 with optimized thresholds, Algorithm 1 with naive thresholds, and MMSE equalization with SIC, with different SNR levels, for the case where each user is allocated a 4-QAM constellation.

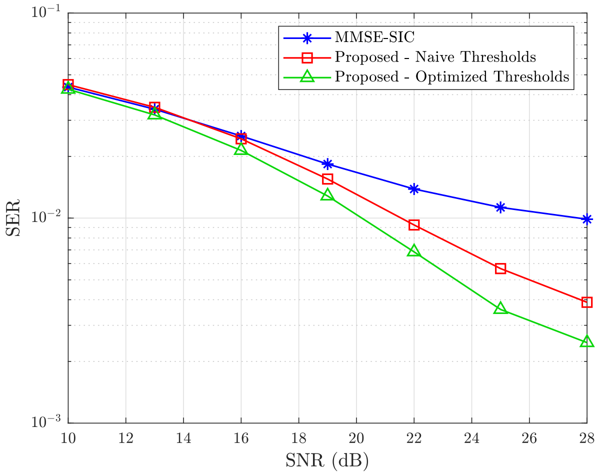

Fig. 5. Comparison of the SER performance of User 1 using Algorithm 1 with optimized thresholds, Algorithm 1 with naive thresholds, and MMSE equalization with SIC, with different SNR levels, for the case where each user is allocated a 16-QAM constellation.

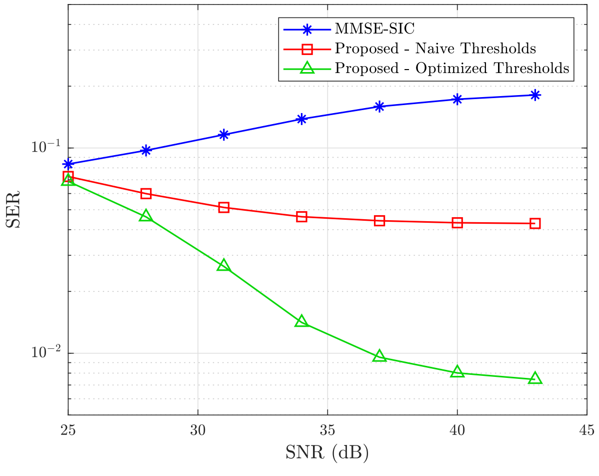

Fig. 6. Comparison of the SER performance of User 2 using Algorithm 1 with optimized thresholds, Algorithm 1 with naive thresholds, and MMSE equalization with SIC, with different SNR levels, for the case where each user is allocated a 16-QAM constellation.

Monte Carlo simulation is used to average the results over 10^5 random channel instances.

A carrier frequency of $f_c = 5.9$ GHz, a transmission bandwidth of 4.95 MHz and a delay-Doppler grid size of $M = 64$ and $N = 16$ are considered. Additionally, we consider a fixed SNR difference of 15 dB between the users, i.e., User 2 has an average SNR that is 15 dB higher than that of User 1. The Tapped Delay Line C (TDL-C) model with a delay spread of 300 ns [29] is used for the channel model. We consider a range of maximum Doppler shifts from 500 Hz to 2500 Hz, which corresponds to velocities of approximately 90 km/h to 450 km/h at a carrier frequency of 5.9 GHz. The Doppler shifts are generated using Jakes' model [30]. For the mLSQR algorithm, a maximum number of iterations of $U = 15$ and a tolerance of $\epsilon = 10^{-2}$ are used, which are commonly used values for LSQR implementation in the related literature [10], [25]. Additionally, the (low-complexity) approximate MSE computation method outlined in subsection IV-C is used in the mLSQR algorithm for all simulations. For Algorithm 1, we consider a maximum number of iterations of $K = 10$ to limit the computational complexity. For power allocation, we use the average-SNR-based fractional transmit power allocation (FTPA) scheme outlined in [16]. The scheme works by considering the average SNR of each user as a fraction of the sum of the SNR of both users. The transmit power of User i is given by:

$$\rho_i = \frac{\mathrm{SNR}_i}{\mathrm{SNR}_1 + \mathrm{SNR}_2}.$$

To compare the low-complexity MSE computation outlined in Section IV-C to the exact method outlined in Section IV-B,

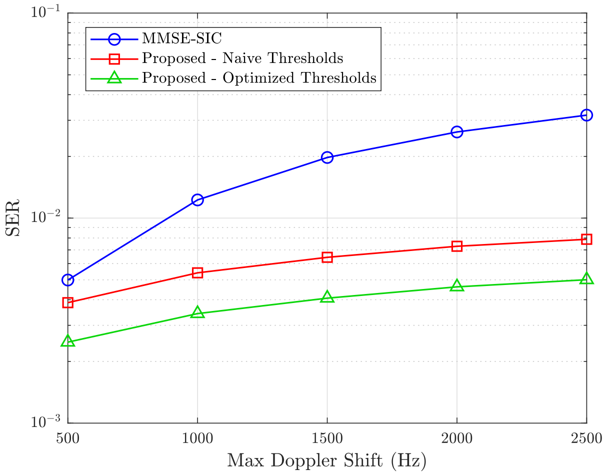

Fig. 7. Comparison of the SER performance of User 1 using Algorithm 1 with optimized thresholds, Algorithm 1 with naive thresholds, and MMSE equalization with SIC, with different maximum Doppler shifts, for the case where each user is allocated a 16-QAM constellation.

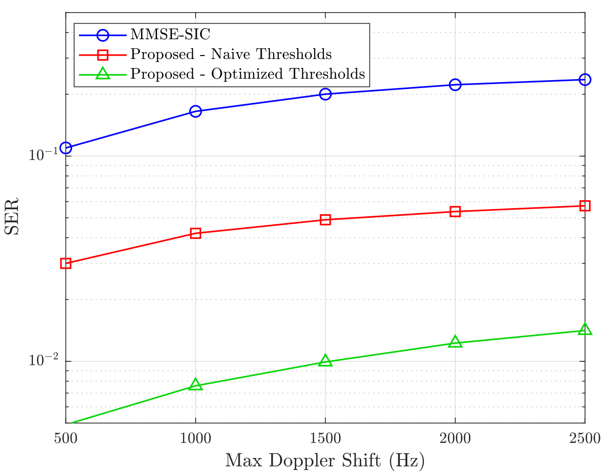

Fig. 8. Comparison of the SER performance of User 2 using Algorithm 1 with optimized thresholds, Algorithm 1 with naive thresholds, and MMSE equalization with SIC, with different maximum Doppler shifts, for the case where each user is allocated a 16-QAM constellation.

we demonstrate the approximation error of the low-complexity method. We define the normalized approximation error as

$$e_\gamma = \frac{1}{MN}\mathbb{E}\left\{\sum_{n=0}^{MN-1} |\gamma_n - \tilde{\gamma}|^2\right\} \tag{50}$$

Figs. 2 shows e_γ at different velocities with an SNR of 15 dB. For this simulation, a small scale example is considered, where $M = N = 4$, due to the computational complexity of the exact MSE computation method. As can be seen in Fig. 2, the approximation error is very small at low velocities, which demonstrates the validity of the low-complexity method when the channel matrix structure is close to BCCB. As expected, the error becomes larger as the velocity increases, as the assumption of a BCCB channel matrix becomes less valid. However, the approximation error is still relatively small and the low-complexity approximate MSE calculation is still useful for choosing the user thresholds.

Fig. 3 shows the symbol error rate (SER) of User 1 using the proposed equalization and detection method compared to the benchmark schemes for different signal-to-noise ratio (SNR) conditions. For these simulations each user's symbols are taken from a 4-QAM constellation, i.e., $A_1 = A_2 = 4$, and the user velocity is fixed at 200 km/h, which equates to a maximum Doppler shift of approximately 1000 Hz. It can be seen from Fig. 3 that for User 1, the proposed method outperforms the MMSE-SIC method, providing an SNR gain more than 2 dB at an SER of 10^{-3}. Additionally, optimizing the RZ detector thresholds provides further performance gains over the naive threshold design benchmark. Since User 1 has a larger power allocation, it is less affected by MUI due to the disparity in the user power levels. Hence, optimizing the the RZ thresholds provides smaller gain than for User 2.

The performance gains of the proposed algorithm primarily come from the proposed iterative detector based on using reliability zones and interference cancellation. At each iteration of Algorithm 1, user symbols are only decoded if they are deemed "reliable" by the RZ detection scheme outlined in Section III-A. The interference cancellation process then removes only these *reliable* symbols from the superimposed symbol vector, producing a superimposed symbol vector with a lower interference level in each iteration with high probability. This in turn aids detection and interference cancellation in the next iteration. The proposed method gains additional performance improvements by optimizing the thresholds which determine the "unreliable zone" for the RZ detection scheme. By optimizing the thresholds to minimize the user MSE at the next iteration, we allow for more interference to be removed at early iterations than the naive conservative threshold design case. This improves detection of the unreliable symbols further and in turn improves the SER performance.

Fig. 4 shows the SER of User 2 using the proposed method compared to the benchmark schemes for different SNR conditions for the 4-QAM case. It can be seen from Fig. 4 that the proposed method significantly outperforms the benchmark schemes. The proposed method with optimized RZ thresholds provides performance gains of many orders of magnitude over the MMSE-SIC scheme and also over the naive threshold design benchmark scheme. This is because the naive threshold design with tight starting thresholds means that fewer User 1 symbols are detected during early iterations and their MUI is still present in the system when the User 2 symbols are being detected. Optimizing the thresholds to minimize User 2 MSE allows for more MUI to be removed at early iterations and improves the accuracy of User 2 symbol detection. Additionally, the proposed method provides significant performance gains over MMSE-SIC which performs very poorly, especially at high SNRs. This is due to the fact that, as the SNR increases, MMSE equalization becomes closer to zero-forcing equalization and the interference is amplified by the inverse matrix involved in the equalization process.

Fig. 5 shows the SER of User 1 using the proposed equalization and detection method compared to the benchmark schemes for different SNR conditions, for the case where each user's symbols are taken from a 16-QAM constellation ($A_1 = A_2 = 16$). For these simulations, the user velocity is fixed at 200 km/h, which equates to a maximum Doppler shift of approximately 1000 Hz. It can be seen from Fig. 3 that for User 1, the proposed method outperforms the MMSE-SIC method, providing an SNR gain of 6 dB. Additionally, optimizing the RZ detector thresholds provides an SNR gain of 2 dB at an SER of 10^{-2} over the naive threshold design benchmark. Fig. 6 shows the SER of User 2 in OTFS-NOMA using the proposed method compared to the benchmark schemes for different SNR conditions for the 16-QAM case. It can be seen that the proposed method outperforms the benchmark the MMSE-SIC scheme for User 2 in the 16-QAM case as well. In addition, optimizing the RZ detector thresholds provides a significant performance increase over the naive threshold design benchmark scheme.

Fig. 7 and Fig. 8 show the SER of User 1 and User 2, respectively, under the proposed equalization and detection method, compared to the benchmark MMSE-SIC scheme, for different values of maximum Doppler shift, for the 16-QAM case. It can be seen that the performance gains of the proposed method over MMSE-SIC actually improves in high Doppler environments, as the performance of MMSE-SIC deteriorates significantly at higher maximum Doppler shifts. This is because as the Doppler shift increases, the channel matrix is more likely to be ill-conditioned; hence, the matrix inversion involved in MMSE equalization may not be robust and can introduce significant equalization error. Additionally, the optimized RZ threshold design offers a significant performance improvement over the naive threshold design for both users. This confirms the benefits of optimizing the RZ thresholds, especially for the user with the lower power allocation.

VII. Conclusion

This paper has presented a novel receiver for downlink OTFS-NOMA. The proposed method uses an iterative process which deploys the LSQR algorithm to equalize the channel, RZ detection to detect symbols from both users within each iteration, and interference cancellation to remove MUI as well as IDI and ISI. The proposed modifications to the LSQR algorithm calculates the post-equalization MSE information needed for optimizing the RZ thresholds. An exact method was presented for computing the MSE as well as a low-complexity approach which takes advantage of the properties of the delay-Doppler channel in OTFS. By optimizing the thresholds, we are able to remove more MUI from the system at early iterations and are therefore able to improve detection performance on subsequent iterations. Numerical results demonstrate the superiority of the proposed method, in terms of SER performance, with respect to an MMSE-SIC benchmark scheme and with respect to a corresponding scheme with naive, pre-determined RZ threshold design.

References

[1] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G Wireless Systems: Vision, Requirements, Challenges, Insights, and Opportunities," *Proceedings of the IEEE*, vol. 109, no. 7, pp. 1166–1199, 2021.

[2] Z. Wei, W. Yuan, S. Li, J. Yuan, G. Bharatula, R. Hadani, and L. Hanzo, "Orthogonal Time-Frequency Space Modulation: A Promising Next-Generation Waveform," *IEEE Wireless Communications*, vol. 28, no. 4, pp. 136–144, 2021.

[3] R. Hadani, S. Rakib, M. Tsatsanis, A. Monk, A. J. Goldsmith, A. F. Molisch, and R. Calderbank, "Orthogonal Time Frequency Space Modulation," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, 2017, pp. 1–6.

[4] S. Tiwari, S. S. Das, and V. Rangamgari, "Low complexity LMMSE Receiver for OTFS," *IEEE Communications Letters*, vol. 23, no. 12, pp. 2205–2209, 2019.

[5] G. D. Surabhi and A. Chockalingam, "Low-Complexity Linear Equalization for OTFS Modulation," *IEEE Communications Letters*, vol. 24, no. 2, pp. 330–334, 2020.

[6] T. Zou, W. Xu, H. Gao, Z. Bie, Z. Feng, and Z. Ding, "Low-Complexity Linear Equalization for OTFS Systems with Rectangular Waveforms," in *IEEE International Conference on Communications Workshops*, 2021, pp. 1–6.

[7] P. Raviteja, K. T. Phan, Y. Hong, and E. Viterbo, "Interference Cancellation and Iterative Detection for Orthogonal Time Frequency Space Modulation," *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6501–6515, 2018.

[8] G. D. Surabhi, M. K. Ramachandran, and A. Chockalingam, "OTFS Modulation with Phase Noise in mmWave Communications," in *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*, 2019, pp. 1–5.

[9] M. Kollengode Ramachandran and A. Chockalingam, "MIMO-OTFS in High-Doppler Fading Channels: Signal Detection and Channel Estimation," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 206–212.

[10] H. Qu, G. Liu, L. Zhang, S. Wen, and M. A. Imran, "Low-Complexity Symbol Detection and Interference Cancellation for OTFS System," *IEEE Transactions on Communications*, vol. 69, no. 3, pp. 1524–1537, 2021.

[11] S. Rakib and R. Hadani, "Multiple access in wireless telecommunications system for high-mobility applications," U.S. Patent 9 722 741 B1, August, 2017.

[12] G. D. Surabhi, R. M. Augustine, and A. Chockalingam, "Multiple Access in the Delay-Doppler Domain using OTFS modulation," 2019.

[13] R. Chong, S. Li, J. Yuan, and D. W. K. Ng, "Achievable Rate Upper-Bounds of Uplink Multiuser OTFS Transmissions," *IEEE Wireless Communications Letters*, vol. 11, no. 4, pp. 791–795, 2022.

[14] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A Survey of Non-Orthogonal Multiple Access for 5G," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2294–2323, 2018.

[15] Z. Ding, R. Schober, P. Fan, and H. Vincent Poor, "OTFS-NOMA: An Efficient Approach for Exploiting Heterogenous User Mobility Profiles," *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7950–7965, 2019.

[16] A. Chatterjee, V. Rangamgari, S. Tiwari, and S. S. Das, "Nonorthogonal Multiple Access With Orthogonal Time–Frequency Space Signal Transmission," *IEEE Systems Journal*, vol. 15, no. 1, pp. 383–394, 2021.

[17] K. Deka, A. Thomas, and S. Sharma, "OTFS-SCMA: A Code-Domain NOMA Approach for Orthogonal Time Frequency Space Modulation," *IEEE Transactions on Communications*, vol. 69, no. 8, pp. 5043–5058, 2021.

[18] H. Wen, W. Yuan, and S. Li, "Downlink OTFS Non-Orthogonal Multiple Access Receiver Design based on Cross-Domain Detection," in *IEEE International Conference on Communications Workshops*, 2022, pp. 928–933.

[19] G. Taubock, M. Hampejs, P. Svac, G. Matz, F. Hlawatsch, and K. Grochenig, "Low-Complexity ICI/ISI Equalization in Doubly Dispersive Multicarrier Systems Using a Decision-Feedback LSQR Algorithm," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 2432–2436, 2011.

[20] S. McWade, M. F. Flanagan, and A. Farhang, "Low-Complexity Equalization and Detection for OTFS-NOMA," 2022. [Online]. Available: https://arxiv.org/abs/2211.07388

[21] A. Farhang, A. RezazadehReyhani, L. E. Doyle, and B. Farhang-Boroujeny, "Low Complexity Modem Structure for OFDM-Based Orthogonal Time Frequency Space Modulation," *IEEE Wireless Communications Letters*, vol. 7, no. 3, pp. 344–347, 2018.

[22] M. Hampejs, P. Svac, G. Taubock, K. Grochenig, F. Hlawatsch, and G. Matz, "Sequential LSQR-based ICI equalization and decision-feedback ISI cancellation in pulse-shaped multicarrier systems," in *IEEE*

10th Workshop on Signal Processing Advances in Wireless Communications, 2009, pp. 1–5.

[23] C. C. Paige and M. A. Saunders, "LSQR: An Algorithm for Sparse Linear Equations and Sparse Least Squares," *ACM Trans. Math. Softw.*, vol. 8, no. 1, p. 43–71, mar 1982.

[24] T. Hrycak, S. Das, G. Matz, and H. G. Feichtinger, "Low Complexity Equalization for Doubly Selective Channels Modeled by a Basis Expansion," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5706–5719, 2010.

[25] H. Qu, G. Liu, Y. Wang, Q. Chen, C. Yi, and J. Peng, "A Time-Domain Approach to Channel Estimation and Equalization for the SC-FDM System," *IEEE Transactions on Broadcasting*, vol. 65, no. 4, pp. 713–726, 2019.

[26] B. Yin, M. Wu, J. R. Cavallaro, and C. Studer, "Conjugate gradient-based soft-output detection and precoding in massive MIMO systems," in *IEEE Global Communications Conference*, 2014, pp. 3696–3701.

[27] Q. He, Y. Hu, and A. Schmeink, "Closed-Form Symbol Error Rate Expressions for Non-Orthogonal Multiple Access Systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 6775–6789, 2019.

[28] R. P. Brent, "An Algorithm with Guaranteed Convergence for Finding a Zero of a Function," *The Computer Journal*, vol. 14, pp. 422–425, 1971.

[29] *3GPP TS 38.901*, 3rd Generation Partnership Project (3GPP), June 2018, v15.0.0.

[30] C. Xiao, Y. Zheng, and N. Beaulieu, "Second-order statistical properties of the WSS Jakes' fading channe simulator," *IEEE Transactions on Communications*, vol. 50, pp. 888 – 891, 07 2002.

Stephen McWade (Member, IEEE) received the Ph.D. degree in electronic engineering from University College Dublin, Dublin, Ireland, in 2023. Prior to this, he received the B.A.I. and M.A.I. degrees in electronic engineering from Trinity College Dublin, Dublin, Ireland, in 2018. He is currently a Research Fellow at the Department of Electronic and Electrical Engineering at Trinity College Dublin. His research interests broadly include wireless communications, multiuser communications and digital signal processing.

Arman Farhang (Senior Member, IEEE) received the Ph.D. degree from Trinity College Dublin, Dublin, Ireland, in 2016.

He was a Research Fellow with CONNECT, Trinity College Dublin, from 2016 to 2018, and an Assistant Professor with Maynooth University, Maynooth, Ireland, and University College Dublin, Dublin, Ireland, from 2018 to 2021. He is currently an Assistant Professor with the Department of Electronic and Electrical Engineering at Trinity College Dublin. He has authored or co-authored over 60 peer-reviewed international journal and conference papers, 4 book chapters, 1 edited book, and he holds 3 patents. Dr Farhang is the director of the NEW WAVE lab at Trinity College Dublin where is leads multiple research projects that are funded by Science Foundation Ireland (SFI). His research interests include wireless communications, digital signal processing for communications, waveform design, multiuser communications, multiantenna, and multicarrier systems.

He is an Associate Editor for the EURASIP Journal on Wireless Communications and Networking. He was a Member of the Organization Committee of the IEEE conference ICC 2020 and the Workshop on OTFS and Delay-Doppler Multicarrier Communications for 6G in ICC 2023. He regularly serves as the TPC in top-tier IEEE conferences and workshops in addition to being an active reviewer of several major IEEE journals.

Mark F. Flanagan (Senior Member, IEEE) received the B.E. and Ph.D. degrees in electronic engineering from University College Dublin (UCD), Dublin, Ireland, in 1998 and 2005, respectively. He is currently a Professor with the School of Electrical and Electronic Engineering, UCD, having been first appointed as SFI Stokes Lecturer in 2008. Prior to this, he held post-doctoral research fellowship positions with the University of Zurich, Switzerland, the University of Bologna, Italy, and the University of Edinburgh, UK. In 2014, he was a Visiting Senior Scientist with the Institute of Communications and Navigation, German Aerospace Center, Munich, under a DLR-DAAD Fellowship. He has published more than 170 papers in peer-reviewed international journals and conferences. His research interests broadly span wireless communications, coding and information theory, and signal processing. He was a recipient of the Stokes Lectureship Award from Science Foundation Ireland (SFI) in 2008 and the Consolidator Laureate Award from the Irish Research Council (IRC) in 2018. He was a recipient of the Best Paper Award at Globecom 2021. He served as TPC Co-Chair for the Communication Theory Symposium at IEEE ICC 2020 and at IEEE GLOBECOM 2022. He is currently serving as TPC Co-Chair for the Wireless Communications Symposium at IEEE ICC 2024. He is also serving as Secretary of the IEEE Radio Communications Society and as Vice-Chair for the Special Interest Group on Reconfigurable Intelligent Surfaces of the Signal Processing and Computing for Communications (SPCC) Technical Committee of the IEEE Communications Society. During the period 2012–2021 he served in the roles of Editor, Senior Editor, and Executive Editor for IEEE Communications Letters. He is currently serving as an Editor for IEEE Transactions on Communications.