A Complexity-Reduced QRD-SIC Detector for Interleaved OTFS

Hua Shao, Haijun Zhang *Senior Member, IEEE*, Huan Zhou, Jianquan Wang, Ning Wang,

Arumgam Nallanathan *Fellow, IEEE*

Abstract

Signal detectors are quite important to attain the diversity of doubly-dispersive wireless channels. Detectors based on message-passing (MP) of factor graphs have been regarded as the way to achieve the near-optimal performance for OTFS. In this paper, by deriving the pattern of the multipath vectorized channel matrix of the orthogonal time frequency space (OTFS) system, it is shown that short girth (i.e. girth-4) may exist in the Tanner graphs, which will degrade the performance of MP detectors, especially with high modulation orders. By introducing interleavers at the transmitter and receiver, the vectorized channel matrix turns out to be a sparse upper block Heisenberg matrix, whose structure is beneficial for the computation of matrix QR decomposition (QRD). Successive interference canceling (SIC) detectors based on QRD and sorted QRD are constructed to eliminate the cross-symbol interference and improve the reliability of the symbol-level channel. Simulation results show that for 4QAM, the QRD-based SIC detectors can achieve about 4dB gain at 10^{-2} over the non-SIC detectors, while the sorted QRD-based SIC detectors can bring an additional 2dB at 10^{-3}, which is only 1dB gap from the MP. For 16QAM, the sorted SIC detectors show superior BER performance than the MP method, and for

This work is supported by China Postdoctoral Science Foundation under grant 2020M680349, the Fundamental Research Funds for the Central Universities (FRF-TP-20-110A1), National Natural Science Foundation of China (62101030).

Hua Shao, Haijun Zhang, Jianquan Wang are with the University of Science and Technology Beijing, Beijing 100083, China (Email: shaohua@ustb.edu.cn, haijunzhang@ieee.org, wangjianquan@ustb.edu.cn).

Huan Zhou is with the College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China. Email: zhouhuan117@gmail.com

Ning Wang is with Henan Joint International Research Laboratory of Intelligent Networking and Data Analysis, School of Information Engineering, Zhengzhou University (Email: ienwang@zzu.edu.cn).

A. Nallanathan is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (Email:nallanathan@ieee.org)

64QAM, the MP detector reaches the error floor while SIC detectors show their excellent performance in all configurations.

Index Terms

QR decomposition, Temporal interleaver, Block Heisenberg matrix, OTFS, Detectors

I. Introduction

Waveform design is the pioneer in the radio access network (RAN) of the 3rd generation partnership project (3GPP). Orthogonal frequency division multiplexing (OFDM) can transform the frequency-selective channel into multiple flat-fading channels by splitting the wideband channel into multiple parallel subchannels. However, this waveform meets huge challenges in the doubly-dispersive channel, i.e. fast-moving scenarios such as high-speed train (HST), V2X, etc. With the rapid changing of the wireless environment, more radio resource needs to be allocated for channel estimation for OFDM, which would degrade the system throughput. Besides, the diversity may be degraded due to the cross-subcarrier interference caused by the corruption of orthogonality due to the multipath Doppler shift.

OTFS waveform has been invented to combat the doubly-dispersive wireless channel[1][2]. At the transmitter, the original information symbols are placed on the delay-Doppler (DD) plane, and an inverse symplectic finite Fourier transformation (ISFFT) is applied to transform the DD domain symbols to the time-frequency domain. Then, the Heisenberg transformation is performed before transmitting the signal. At the receiver side, the corresponding inverse transformations are performed to obtain the delay-Doppler signal. Since each symbol on the delay-Doppler plane is spread over the whole time-frequency plane, OTFS has full diversity. Evidence shows that OTFS has superior performance than the OFDM in fast-moving scenarios.

OTFS can be implemented by adding pre-and post-processing blocks to the traditional OFDM system[3], i.e. the ISFFT and symplectic finite Fourier transformation (SFFT). Detecting is performed on the delay-doppler domain, which is block-based. Detecting and equalizing over the signal block are quite critical to attaining the full diversity of OTFS. Since the number of paths in fast-moving scenarios may be limited, the support of the DD domain channel is sparse, which could simplify the channel estimation and signal detection. As the delay-Doppler channel can be expressed by a sparse matrix, message passing (MP)-based detectors are designed by transforming the sparse channel matrix into factor graphs[4]. The belief is exchanged between

the variable nodes and checkout nodes from iteration to iteration, and the final result is achieved when the stopping criteria is met. MP methods have been regarded as the near-optimal solution for a factor graph without cycles[5].

In this paper, the vectorized channel model for OTFS is analyzed, and it is shown that there is short girth in the Tanner graph corresponding to the input-output relation. The existence of such short girth may lead to the MP method converging to a local optimum and severely degrade the final performance, as in regular LDPC code construction. The simulation also reveals the performance degradation with higher modulation orders caused by such a structure.

Besides, by introducing interleavers at the transmitter and receiver, it is shown that the delay-Doppler channel is a sparse upper block Heisenberg matrix, which can be QR decomposed (QRD) fastly. Based on the QRD of the channel matrix, both non-sorted and sorted SIC detectors are constructed in this paper, evidence shows their superior performance over the MP method, especially with high modulation order.

The contributions of this paper are three folds:

- By analyzing the structure of the vectorized channel of the delay-Doppler domain, it is shown that short girth exists in the Tanner graph, which would severely degrade the decoding performance of the MP-based method.
- Interleavers are introduced into the transmitting-receiving chain without any other computation burden. As a result, the delay-doppler channel matrix has the form of a sparse upper block Hessenberg matrix, and all the non-zero elements under the main diagonal will locate in a sub-diagonal block with a fixed size. The sparse block Heisenberg matrix can be QR decomposed fastly. Then, the SIC detectors can be implemented based on the fixed special channel structure.
- Different configurations including low (4QAM), median (16QAM), and high (64QAM) modulation order with different frame sizes are evaluated to compare the performance. The results show that the QRD-based SIC detectors have good performance in both the BER and diversity in all configurations, while the performance of the MP detector will degrade or even reaches its error floor in high modulation order due to the existence of short girth.

The rest of the paper is structured as follows. The related works are summarized in section II, the system model and the vectorized channel model are presented in Section III. The existence of short girth and the QRD-based SIC detectors are given in Section IV. The simulation and performance comparison are shown in Section V. Finally Section VI gives the conclusion.

Notations: Throughout, $\mathbf{x}$ and $\mathbf{X}$ are used for denoting vectors and matrix respectively, $\mathbf{X}^H$, $\mathbf{X}^T$ and $\mathbf{X}^*$ denote the hermitian transpose, transpose and conjugates of $\mathbf{X}$ respectively. $||.||$ denotes the ℓ_2-norm of vectors or matrix over complex or real space. $vec(\mathbf{T})$ is the vector by stacking the columns of $\mathbf{T}$. $\otimes$ denotes the kronecker product. $\mathbb{C}, \mathbb{R}$ denotes the complex and real domain respectively.

II. Related Work

OTFS has a close relationship with the existing waveforms. By adding precoding and post-processing units of OFDM, the ISFFT and SFFT of OTFS can be implemented in a low-complexity way[6]. In [7], the work derived an upper bound on the PAPR of the OTFS signal and showed that the maximum PAPR grows linearly with the number of Doppler bins, which is different from OFDM [8]. In [9], a simple matrix form of the input-output relation in the DD domain for rectangle pulse-shaping waveforms was derived. In [10], the author showed that in static multipath channels, the system structure of OTFS is equivalent to the asymmetric orthogonal frequency division multiplexing (A-OFDM), and an MP detection method was proposed to show performance gain over zero-forcing (ZF) and minimum mean square error (MMSE). The work In [11] presented a comparison between OTFS and OFDM in terms of achievable communication rate, where the channel sparsity, guard interval, and cyclic prefix are considered. The work in [12] proposed a modulation framework which is a generalization of OTFS, and the low-complexity waveform based on Walsh-Hadamard transform (WHT) exhibits the same frame error rate (FER) as OTFS.

Channel estimation of OTFS can exploit the sparse nature of multipath on the DD plane. In [13], the authors proposed an embedded pilot-aided channel estimation scheme for OTFS, where the pilot, guard, and data symbols in the delay-Doppler plane are ranged suitably to avoid interference between pilot and data symbols at the receiver. The work in [14] showed that the OTFS multiple-input multiple-output (MIMO) channel exhibits 3D-structured sparsity, and a channel estimation technique based on 3D-structured orthogonal matching pursuit (OMP) algorithm is proposed. The work in [15] exploited the angle, delay, and Doppler reciprocity between the uplink and the downlink and reconstructed the angles, the delays, and the doppler frequencies for the downlink massive channels at the base station. The work in [16] proposed a block-wise OTFS receiver by exploiting the structure and characteristics of the OTFS transmission matrix, which can realize fast convergence by leveraging the sparsity. In [17], a sparse channel state

information (CSI) estimation model was proposed for reducing the pilot overhead of OTFS modulation aided MIMO system, which completely avoids placing multiple DD-domain guard intervals corresponding to each transmit antenna within the same OTFS frame. The work in [18] investigated the impacts of transmitter and receiver windows on the performance of OTFS modulation and proposed window designs to improve the OTFS channel estimation and data detection performance. The fractional doppler effect is also reduced with the proposed window designs.

Equalizing and detecting play important roles in attaining the diversity of OTFS. In [4], the authors derived the explicit input-output relation describing OTFS modulation and demodulation considering both the ideal and rectangle windows. An MP detector is also proposed based on the sparse channel matrix. The work in [19] proved that the asymptotic diversity order of the OTFS is one. While in the finite SNR regime, the potential for a higher-order diversity is witnessed before the diversity one regime takes over. The work in [20] proposed a low complexity linear MMSE receiver that exploits the sparsity and quasi-banded structure of matrices. It results in a log-linear order of complexity without any performance degradation of bit error rate (BER). The work in [21] studied the diversity of OTFS assuming rectangular waveforms and a DD channel with two paths. The concept of effective diversity was introduced which is more significant than the existing standard metric. The work in [22] proposed a low complexity equalization approach that gives exact MMSE and ZF solutions with just $\mathcal{O}(MNlogMN)$ complexity. The work in [23] proposed a variational Bayes approach as an approximation of the optimal maximum a posteriori probability (MAP) detection which is guaranteed to converge to the global optimum. The work in [24] proposed a low complexity iterative decision feedback equalizer (DFE) by using the vector input-output relation. In [25], a two-stage equalizer enjoying low implementation complexity was proposed, where a sliding window-assisted equalizer is used to restrict the inter-carrier interference (ICI) in the time-frequency domain and a DD domain equalizer is used to mitigate the residual inter-symbol interference (ISI). The work in [26] designed detectors that enjoy the structure of the channel matrix and allow efficient implementation by leveraging the unitary AMP (UAMP). In [27], a cross-domain iterative detection algorithm was proposed to enhance the error performance of OTFS modulation, where both the time domain and DD domain are applied. The work in [28] devised a fractionally spaced sampling receiver in which the sampling rate is an integer multiple of the symbol rate, which demonstrates stronger performance than traditional receivers.

OTFS has also been in conjunction with other techniques to meet the requirements of new scenarios[29]. In [30], A new non-orthogonal multiple-access (NOMA) transmission protocol that incorporates OTFS modulation was proposed. The work in [31] proposed a new path division multiple access (PDMA) by scheduling the users in the non-overlap angle-delay-Doppler domain. In [32], the authors considered a scenario where the integrated (radar) sensing and communication (ISAC) technique for assisting OTFS transmission in both uplink and downlink vehicular communication systems is deployed.

III. System Model

An interleaved uncoded OTFS modulation system is considered in this paper, the main modules are illustrated in Fig.1. After the bits are mapped to the constellation points, they are placed on the delay-Doppler domain, and an ISFFT is performed on the 2D data symbols. The Heisenberg transform is performed on the 2D data symbols to generate the time-domain samples. Different from the conventional OTFS, an interleaver is inserted before the ISFFT, which is used to reform the structure of the end-to-end channel. After the de-interleaver which is used to align the data, a cyclic prefix (CP) is inserted at the end of each data block, as stated in [9]. It is noted that this is different from OFDM, where the CP is inserted at the beginning of each temporal symbol. Then the signals are transmitted through a linear time-varying (LTV) channel. At the receiver, after the CP removing, the same interleaver is inserted to reform the receiving channel structure. The Wigner transformation and the succeeding SFFT are performed in a sequel to transform the receive data to the delay-Doppler domain. Finally, QRD-based detecting is performed at the delay-doppler domain with the aid of the CSI.

A. Conventional OTFS modulation

In this paper, an OTFS system with M subcarriers and N temporal symbols is considered. At the transmitter, the data to be transmitted is firstly mapped to a contellation alphabet of size Q, i.e. $\mathcal{A} = \{a_1, a_2, \cdots, a_Q\}$. Then, MN mapped symbols are placed on the $M \times N$ delay-doppler domain. The $M \times N$ data symbols of the delay-doppler domain are transformed to the time-frequency domain with the ISFFT:

$$\mathbf{X}_{m,n}^{tf} = \frac{1}{\sqrt{NM}} \sum_{l=0}^{M-1} \sum_{k=0}^{N-1} \mathbf{X}_{l,k}^{dd} e^{j2\pi(\frac{nk}{N} - \frac{ml}{M})}, \tag{1}$$

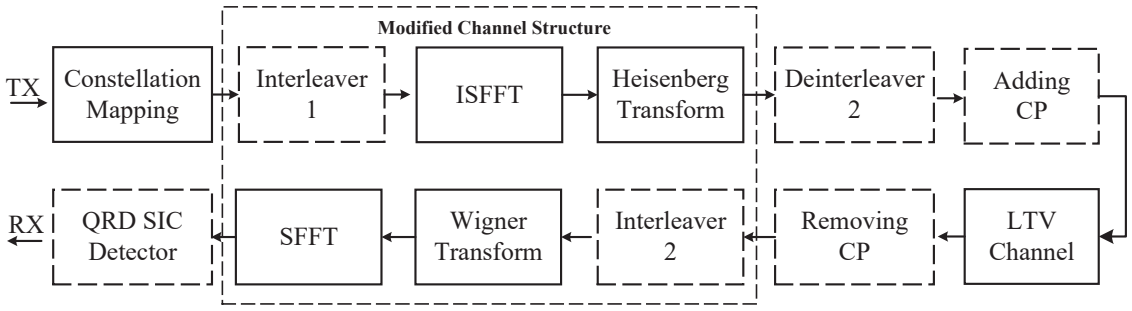

Fig. 1: An uncoded interleaved OTFS modulation.The interleaver 1 and interleaver 2 are used to transform $\mathbf{F}_N^H \otimes \mathbf{I}_M$ to $\mathbf{I}_M \otimes \mathbf{F}_N^H$, and the vectorized channel would be an upper block hessenberg matrix. Deinterleaver 2 is inserted to align the data samples. QRD-based SIC detecting is performed after the SFFT at the receiver.

where $\mathbf{X}_{l,k}^{dd} \in \mathcal{A}$ denotes the data symbol of the DD domain at index (l,k), k and l are the doppler and delay axis respectively, and $\mathbf{X}_{m,n}^{tf}$ is the TF domain signal at index (m,n) where $n = 0, 1, \cdots, N-1$ and $m = 0, 1, \cdots, M-1$. Then, an Heisenberg transform is performed on $\mathbf{X}_{m,n}^{tf}$ to transform the samples to time domain:

$$s(t) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \mathbf{X}_{m,n}^{tf} g_{tx}(t - nT) e^{j2\pi m \Delta f (t-nT)}, \tag{2}$$

where $g_{tx}(t)$ is the window function, T and Δf are the temporal duration of the temporal symbol and subcarrier bandwidth respectively.

After the Heisenberg transform, CP is appended to the data block of size MN, as in [9]. Only considering one-symbol CP can improve the spectrum efficieny. At the receiver, the CP is removed. CP appending is equivalent to CP inserting such as in OFDM or GFDM. The vectorized channel would be simpler for CP appending, which is easier to be implemented[17].

Then, $s_{cp}(t)$ is transmitted through a doubly-dispersive channel composed of P paths, which can be expressed as:

$$h(\tau, v) = \sum_{i=0}^{P-1} \delta(\tau - \tau_i) \delta(v - v_i), \tag{3}$$

where τ_i and v_i are the delay and doppler of the i-th path respectively.

At the receiver, the received signal can be expressed as:

$$r(t) = \int \int h(\tau, v) s_{cp}(t - \tau) e^{j2\pi v(t-\tau)} d\tau dv \tag{4}$$

The time-frequency domain signal can be retrieved by sampling the cross-ambiguity function of the receive signal $r(t)$ and a receiving window $g_{rx}(t)$:

$$\mathbf{Y}_{m,n}^{tf} = \int g_{rx}^{*}(t'-t)r(t')e^{-2\pi f(t'-t)}dt'|_{f=m\Delta f, n=nT} \tag{5}$$

At last, the SFFT is performed on $\mathbf{Y}_{m,n}^{tf}$ to generate the delay-doppler signal at the receiver:

$$\mathbf{Y}_{l,k}^{dd} = \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \mathbf{Y}_{m,n}^{tf} e^{j2\pi(\frac{ml}{M} - \frac{nk}{N})} \tag{6}$$

According to Equation (1) and (6), an advantage of OTFS is that for each symbol $\mathbf{X}_{l,k}^{dd}$ on the DD domain, it spreads onto the whole $M \times N$ time-frequency resource plane and can potentially achieve the full time-frequency diversity.

B. Vectorized Channel Model

The ISFFT process can be implemented by combining an IDFT along the doppler domain and a DFT along the delay domain:

$$\mathbf{X}^{tf} = \mathbf{F}_M \mathbf{X}^{dd} \mathbf{F}_N^H, \tag{7}$$

where $\mathbf{F}_M$ is the M-point DFT matrix. It is noted that the DFT matrix is unitary and a special case of the vandermonde matrix, which implies $\mathbf{F}_M^{-1} = \mathbf{F}_M^H$. The Heisenberg transform can be expressed as the IDFT along each of the columns of $\mathbf{X}^{tf}$:

$$\mathbf{X}^{t} = (\mathbf{F}_M^H \mathbf{F}_M)\mathbf{X}^{dd}\mathbf{F}_N^H = \mathbf{X}^{dd}\mathbf{F}_N^H \in \mathbb{C}^{M\times N} \tag{8}$$

Considering the rectangle window of the transmitting filter, $\mathbf{X}^t$ can be vectorized by stacking each colum:

$$\mathbf{x}^t = vec(\mathbf{X}^t) = (\mathbf{F}_N^H \otimes \mathbf{I}_M) vec(\mathbf{X}^{dd}) \in \mathbb{C}^{MN\times 1}, \tag{9}$$

where $\mathbf{I}_M$ denotes the identity matrix of $M \times M$ and $(\mathbf{F}_N^H)^T = \mathbf{F}_N^H$.

With interger doppler and delay taps, for the LTV channel composed of P paths, the channel matrix can be expressed as[9]:

$$\mathbf{H}_o = \sum_{i=0}^{P-1} h_i \mathbf{\Gamma}_{MN}^{\tau_i} \mathbf{\Delta}_{MN}^{v_i}, \tag{10}$$

in which τ_i and v_i are the powers of the permutation matrix $\mathbf{\Gamma}_{MN}$ and diagonal matrix $\mathbf{\Delta}_{MN}$ of size $MN \times MN$ respectively, and:

$$\begin{aligned}\mathbf{\Gamma}_{MN} &= \begin{bmatrix} 0 & 0 & \cdots & 1 \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}_{MN}, \\ \mathbf{\Delta}_{MN} &= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & e^{\frac{j2\pi}{MN}} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & e^{\frac{j2\pi(MN-1)}{MN}} \end{bmatrix}_{MN}\end{aligned} \tag{11}$$

Accordingly, at the reciever the recieved signal after the combined FFT and SFFT can be expressed as:

$$\mathbf{y}^{dd} = \mathbf{H}\mathbf{x}^{dd} + \tilde{\mathbf{n}}, \tag{12}$$

where

$$\begin{aligned}\mathbf{x}^{dd} =& vec(\mathbf{X}^{dd}) \\ \mathbf{H} =& (\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{H}_o(\mathbf{F}_N^H \otimes \mathbf{I}_M), \\ \tilde{\mathbf{n}} =& (\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{n}\end{aligned} \tag{13}$$

is the delay-doppler channel.

It is noted that $\mathbf{H}_o$ is the TF domain channel. The main difference between $\mathbf{H}_o$ and $\mathbf{H}$ is the ISFFT and SFFT together with the rectangle windowing effect at the transmitter and receiver. Besides, since $\mathbf{F}_N \otimes \mathbf{I}_M$ is unitary, it does not affect the noise term and $\tilde{\mathbf{n}}$ is still gaussian.

For $\mathbf{H}_o$, if the doppler frequency offset $v_i = 0$, it would be a circulant matrix since $\mathbf{\Gamma}_{MN}$ is a permutation matrix. With $v_i \neq 0$, $\mathbf{H}_o$ will lost the circulant property but the pattern (i.e. the absolute value) still holds because $\mathbf{\Delta}_{MN}$ is an unitary diagonal matrix. Furthermore, since $\tau_i \leq M$, at most M sub-diagonal elements of $\mathbf{H}_o$ would be non-zero. In all, $\mathbf{H}_o$ is an upper Heisenberg matrix with at most M non-zero sub-diagonal elements.

For $\mathbf{H}$, since the Kronecker product $\otimes$ is a left direct product, even though $\mathbf{H}_o$ is a sparse and Hessenberg matrix, the delay-Doppler channel matrix $\mathbf{H}$ lost its Heisenberg properties.

C. Girth of the DD domain channel

Since $\mathbf{H}$ is the sum of P complex matrix, and each complex matrix denotes the complex gain of the corresponding path, the final structure of $\mathbf{H}$ is mainly determined by the permutation matrix and the corresponding delay taps.

Without loss of generality, only one path channel matrix denoted as $\mathbf{H}_i$ with delay τ_i and doppler v_i is considered, and denote $\mathbf{\Pi}_i = \mathbf{\Gamma}_{MN}^{\tau_i}$, $\mathbf{C}_i = \mathbf{\Delta}_{MN}^{v_i}$, and

$$\mathbf{H}_i = h_i \cdot \underbrace{\begin{bmatrix} f_{11}\mathbf{I}_M & \cdots & f_{1N}\mathbf{I}_M \\ f_{21}\mathbf{I}_M & \cdots & f_{2N}\mathbf{I}_M \\ \vdots & \ddots & \vdots \\ f_{N1}\mathbf{I}_M & \cdots & f_{NN}\mathbf{I}_M \end{bmatrix}}_{\mathbf{H}_1} \mathbf{\Pi}_i \, \mathbf{C}_i \underbrace{\begin{bmatrix} f_{11}^*\mathbf{I}_M & \cdots & f_{N1}^*\mathbf{I}_M \\ f_{12}^*\mathbf{I}_M & \cdots & f_{N2}^*\mathbf{I}_M \\ \vdots & \ddots & \vdots \\ f_{1n}^*\mathbf{I}_M & \cdots & f_{NN}^*\mathbf{I}_M \end{bmatrix}}_{\mathbf{H}_2}, \tag{14}$$

where $f_{p,q} \in \mathbb{C}$ is the (p,q)-th element of the DFT matrix $\mathbf{F}_N$, $f_{p,q} = f_{q,p}$.

Now considering the $\mathbf{H}_1 = (\mathbf{F}_N \otimes \mathbf{I}_M)\mathbf{\Pi}_i$ part of $\mathbf{H}_i$, since $\mathbf{\Pi}_i$ is the column permutation of $\mathbf{F}_N \otimes \mathbf{I}_M$, and for each row-block, it is composed of the same pattern (i.e. the non-zero locations, N diagonal matrix), then the effect of the permutation $\mathbf{\Pi}_i$ is equivalent to the block-wise permutation of each block diagonal matrix with $\tau_i < M$. Finally, the pattern of $\mathbf{H}_1$ should satisfy:

$$\mathbf{H}_1^p \triangleq \begin{bmatrix} \mathbf{\Pi}_M^i & \cdots & \mathbf{\Pi}_M^i \\ \mathbf{\Pi}_M^i & \cdots & \mathbf{\Pi}_M^i \\ \vdots & \ddots & \vdots \\ \mathbf{\Pi}_M^i & \cdots & \mathbf{\Pi}_M^i \end{bmatrix}_{MN}, \tag{15}$$

where $\mathbf{\Pi}_M^i$ is the permutation matrix of size M, the permutation step is determined by the delay taps.

For $\mathbf{H}_2 = \mathbf{C}_i(\mathbf{F}_N^H \otimes \mathbf{I}_M)$, with $\mathbf{C}_i(l) = e^{\frac{2\pi j v_i l}{MN}}$ and $f_{p,q}^* = \frac{1}{\sqrt{N}} e^{\frac{2\pi jpq}{N}}$, each column-block i of $\mathbf{H}_2$ can be obtain:

$$\mathbf{H}_2(:, M_i) = \begin{bmatrix} 1 & & & \\ & e^{\frac{j2\pi v_i}{MN}} & & \\ & & \ddots & \\ & & & e^{\frac{j2\pi(MN-1)v_i}{MN}} \end{bmatrix} \begin{bmatrix} f_{0,i}^*\mathbf{I}_M \\ f_{1,i}^*\mathbf{I}_M \\ \vdots \\ f_{N-1,i}^*\mathbf{I}_M \end{bmatrix} \tag{16}$$

For each column of Eq.(16), the m-th non-zero element can be expressed as:

$$\begin{aligned}\mathbf{H}_2(pM+m,m) &= \frac{1}{\sqrt{N}} e^{\frac{j2\pi v_i(m+pM)}{MN}} e^{\frac{j2\pi pi}{N}} \\ &= \frac{1}{\sqrt{N}} e^{\frac{j2\pi p(v_i+i))}{N}} e^{\frac{j2\pi v_i m}{MN}} \\ &= f^*_{p,v_i+i} e^{\frac{j2\pi v_i m}{MN}},\end{aligned} \tag{17}$$

Eq.17 shows that the multiplication of $\mathbf{C}_i$ results in the cyclic shift (v_i) of blocks of $\mathbf{F}_N^H \otimes \mathbf{I}_M$, with an extra coefficient $e^{\frac{j2\pi v_i m}{MN}}$ per column-block.

The final pattern of the channel matrix $\mathbf{H}_i$ has the following structure:

$$\mathbf{H}_i^p \triangleq \begin{bmatrix} \mathbf{\Pi}_M^i & \cdots & \mathbf{\Pi}_M^i \\ \vdots & \ddots & \vdots \\ \mathbf{\Pi}_M^i & \cdots & \mathbf{\Pi}_M^i \end{bmatrix}_{MN} \mathbf{\Pi}_N^i \quad , \tag{18}$$

where $\mathbf{\Pi}_N = \mathbf{\Gamma_N}^{v_i} \otimes \mathbf{I}_M$ is the column-block permutation introduced by $\mathbf{C}_i$.

From Eq.18, the pattern of $\mathbf{H}_i^p$ can be regarded as the parity check matrix of a quasi-cyclic low-density parity check (QC-LDPC), which is constructed with a base matrix being $\mathbf{\Gamma_N}^{v_i}$, and the power factor of the non-zero positions being τ_i.

Furthermore, since $\mathbf{H}$ is the sum of multiple path, its pattern has the form:

$$\mathbf{H}^p \triangleq \begin{bmatrix} k_{11}\sum_i \mathbf{\Pi}_M^i & \cdots & k_{1n}\sum_i \mathbf{\Pi}_M^i \\ k_{21}\sum_i \mathbf{\Pi}_M^i & \cdots & k_{2n}\sum_i \mathbf{\Pi}_M^i \\ \vdots & \ddots & \vdots \\ k_{n1}\sum_i \mathbf{\Pi}_M^i & \cdots & k_{nn}\sum_i \mathbf{\Pi}_M^i \end{bmatrix}, \tag{19}$$

where $k_{i,j} \in \{0,1\}$ is determined by the multipath Doppler and delay taps. From Eq. (19), it can be seen that the additional multipath would not prolong the overall girth of $\mathbf{H}^p$, but may further decrease it since more non-zeros are added into the parity-check matrix.

As an example, if $t \geq 3$ taps are consective in the delay domain, according to Eq.(19):

$$\sum_i \mathbf{\Pi}_M^i = \begin{matrix} \tau_0 \\ \tau_1 \\ \tau_2 \\ \vdots \end{matrix} \begin{bmatrix} 1 & 0 & \cdots & 1 & 1 \\ 1 & 1 & \ddots & \ddots & 1 \\ 1 & 1 & \ddots & \ddots & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}, \tag{20}$$

where τ_i denotes the permutation item of delay-i. The dashed line in Eq.(20) forms a girth-4.

With the potential short girth, the MP-based method may converge to a local optimal. In fact, it was shown that QC-LDPC codes cannot have a Tanner graph representation with a girth larger

than **12** [33]. A necessary condition to have $g > 6$ in the single-path Tanner graph representation of the (N, N)-regular QC-LDPC code is $M > N$, and a necessary condition to have $g > 8$ in the Tanner graph representation of a (N, N)-regular QC-LDPC code is $M > (N-1)^2$. Increasing the delay bins (M) can benefit the performance of MP decoding.

D. Temporal interleaved OTFS channel

From the previous section, it can be known that the time-frequency channel $\mathbf{H}_o$ has the following form:

$$\mathbf{H}_o = \tau_m \begin{array}{c} \begin{matrix} 0 \\ 1 \\ \vdots \\ \\ \\ \\ \\ \end{matrix} \\ \begin{bmatrix} a_{1,1} & 0 & \cdots & a_{3,p-1} & a_{2,p} \\ a_{2,1} & a_{1,2} & \ddots & \ddots & a_{3,p} \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ a_{m,1} & a_{m-1,2} & \ddots & \ddots & \ddots \\ 0 & a_{m,2} & \ddots & \ddots & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & 0 & \cdots & a_{2,p-1} & a_{1,p} \end{bmatrix}_{MN} \end{array}, \tag{21}$$

where $0, 1, \cdots, \tau_m$ denote the delay taps, and $a_{i,j}$ is the complex gain. $\tau_m < M \ll MN$, $\mathbf{H}_o$ is a sparse upper Heisenberg matrix with compact non-zeros under the main diagonal line. But the SFFT and ISFFT procedures, i.e. $\mathbf{F}_N \otimes \mathbf{I}_M$ and $\mathbf{F}_N^H \otimes \mathbf{I}_M$, permutate the non-zero elements to the positions which is far from the main diagonal line, as illustraed in Equation (19).

However, it is noted that the inverse version, i.e. $\mathbf{I}_M \otimes \mathbf{F}_N$ is a block diagonal matrix. If the left direct product can be converted to the right direct product, then the overall structure of the channel matrix can be simplified. It motivates that an interleaver/deinterleaver pair can be introduced to generate a more compact sparse $\mathbf{H}$.

Proposition: The DD domain channel $\mathbf{H}$ will be $(\mathbf{I}_M \otimes \mathbf{F}_N)\mathbf{H}_o(\mathbf{I}_M \otimes \mathbf{F}_N^H)$ with interleavers inserted into the modified OTFS system as in Fig.1.

Proof: Let's consider the difference between $\mathbf{I}_M \otimes \mathbf{F}_N$ and $\mathbf{F}_N \otimes \mathbf{I}_M$, where

$$\mathbf{F}_N \otimes \mathbf{I}_M = \begin{bmatrix} f_{11}\mathbf{I}_M & \cdots & f_{1N}\mathbf{I}_M \\ \vdots & \ddots & \vdots \\ f_{N1}\mathbf{I}_M & \cdots & f_{NN}\mathbf{I}_M \end{bmatrix} \tag{22a}$$

$$\mathbf{I}_M \otimes \mathbf{F}_N = \begin{bmatrix} \mathbf{F}_N & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{F}_N \end{bmatrix} \tag{22b}$$

For the two rows with the same index (e.g. the i-th row) in Eq.(22a) and Eq.(22b), the non-zero elements have the same values, but with different column indexes. In fact, to keep y_i unchanged after replacing $\mathbf{F}_N \otimes \mathbf{I}_M$ by $\mathbf{I}_M \otimes \mathbf{F}_N$, a permutation of the corresponding input elements should be performed. For th $MN \times 1$ vector output from $\mathbf{H}_o$, **the permutation is just a rectangle interleaver with** $M \times N$.

The same interleaver is inserted before the ISFFT module to transform the $\mathbf{F}_N^H \otimes \mathbf{I}_M$ into $\mathbf{I}_M \otimes \mathbf{F}_N^H$. Additionally, it is noted that a corresponding deinterleaver needs to be added before $\mathbf{H}_o$ to align the data, as illustrated in Fig.1. □

In fact, by adding the interleaving/deinterleaving procedures, the OTFS can be regarded as a configuration of the generalized frequency division multiplexing (GFDM) waveform[34]. However, it is noted that all the modules are implemented with the OTFS framework, and the CP configurations in this framework are different from GFDM.

With the introduced interleavers, the DD channel becomes:

$$\begin{aligned} \mathbf{H} &= (\mathbf{I}_M \otimes \mathbf{F}_N)\mathbf{H}_o(\mathbf{I}_M \otimes \mathbf{F}_N^H) \\ &= \begin{bmatrix} \mathbf{F}_N & & \\ & \ddots & \\ & & \mathbf{F}_N \end{bmatrix} \mathbf{H}_o \begin{bmatrix} \mathbf{F}_N^H & & \\ & \ddots & \\ & & \mathbf{F}_N^H \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{H}_1 & 0 & \cdots & \mathbf{B}''_{M-1} & \mathbf{B}_M \\ \mathbf{B}_1 & \mathbf{H}_2 & \ddots & 0 & 0 \\ 0 & \mathbf{B}_2 & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & \mathbf{H}_{M-1} & 0 \\ 0 & 0 & \cdots & \mathbf{B}'_{M-1} & \mathbf{H}_M \end{bmatrix}, \end{aligned} \tag{23}$$

Since $\mathbf{H}_o$ is a Heisenberg matrix with at most M sub-diagonal elements per column, both $\mathbf{I}_M \otimes \mathbf{F}_N$ and $\mathbf{I}_M \otimes \mathbf{F}_N^H$ are block diagonal matrix with block size $N \times N$, then $\mathbf{H}$ is a sparse

block Heisenberg matrix, with only one block sub-diagonal items being compact and close to the main diagonal line. The sub-diagonal blocks $\mathbf{B}_i$ of $\mathbf{H}$ have a fixed size with $K \times N$, where K is determined by the maximum delay spread:

$$K = \begin{cases} N, & \tau_{max} \leq N \\ 2N, & \tau_{max} > N \end{cases}. \tag{24}$$

It is noted that $\mathbf{B}_{M-1} = \begin{bmatrix} \mathbf{B}'_{M-1} \\ \mathbf{B}''_{M-1} \end{bmatrix} \in \mathbb{C}^{K \times N}$ in Euqation (23). It is noted that the interleavers does not imporove the girth of the factor graph, but the original vectorized channel $\mathbf{H}$ becomes a sparse upper block heisenberg matrix.

$\mathbf{H}$ can be easily QR decomposed with Gram-Schmidt or Givens or Householder transformations by eliminating the non-zeros elements $\mathbf{B}_i$ below the main diagonal. For the block Heisenberg pattern of $\mathbf{H}$ with the same fixed block size $\mathbf{B}_i$, there are a lot of fast algorithms based on such properties[35], which is beyond the scope of this paper.

IV. QRD-based SIC detector for modified OTFS

In this section, based on the previous modified channel structure, QRD-SIC detectors including ZF-SIC and MMSE-SIC are represented utilizing the sparse block Heisenberg properties of the channel matrix, and the complexity comparison is analyzed.

A. QRD-based SIC detectors

The procedures of ZF-SIC and MMSE-SIC detectors are similar, i.e. 1) Decompose the channel matrix into a unitary matrix $\mathbf{Q}$ and upper triangle matrix $\mathbf{R}$. 2) Successive interference cancellation based on the triangle structure of $\mathbf{R}$. MMSE-SIC differs from the channel matrix which is augmented with a regularized block.

- ZF-SIC: The channel matrix is directly decomposed into a $MN \times MN$ unitary matrix and a upper triangle matrix:

$$\mathbf{H} = \mathbf{QR}, \tag{25}$$

with $\mathbf{Q}, \mathbf{R} \in \mathbb{C}^{MN \times MN}$, then:

$$\mathbf{y}_{sic} = \mathbf{Q}^H \mathbf{y}^{dd} = \mathbf{R}\mathbf{x}^{dd} + \tilde{\mathbf{n}}, \tag{26}$$

where $\tilde{\mathbf{n}} = \mathbf{Q}\mathbf{n}$ is gaussian due to $\mathbf{Q}$ is unitary.

- MMSE-SIC: QRD operates on an extended channel matrix with a regularized block, which is denoted:

$$\tilde{\mathbf{H}} = \begin{bmatrix} \mathbf{H} \\ \alpha \mathbf{I}_{MN} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \end{bmatrix} \mathbf{R}, \tag{27}$$

where α is a small regularized factor, $\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{R} \in \mathbb{C}^{MN \times MN}$ and $\mathbf{Q}_1^H \mathbf{H} + \alpha \mathbf{Q}_2^H = \mathbf{R}$, then,

$$\mathbf{y}_{sic} = \mathbf{Q}_1^H \mathbf{y}^{dd} = \mathbf{R}\mathbf{x}^{dd} + \tilde{n}, \tag{28}$$

where $\tilde{\mathbf{n}} = -\alpha \mathbf{Q}_2^H \mathbf{x}^{dd} + \mathbf{Q}_1^H \mathbf{n}$. It is noted that generally $\mathbf{Q}_1 \in \mathbb{C}^{MN \times MN}$ is not unitary due to $[\mathbf{Q}_1^H, \mathbf{Q}_2^H]^H$ is unitary, and $\tilde{\mathbf{n}}$ is coloured noise.

Since $\mathbf{R}$ is an upper triangle matrix, the SIC detector now detects the signal in reverse order, i.e. the MN-th element of $\mathbf{x}^{dd}$:

$$\mathbf{x}^{dd}(MN) = \lceil \frac{\mathbf{y}_{sic}(MN)}{\mathbf{R}(MN, MN)} \rfloor_o, \tag{29}$$

For the rest $MN - 1$ elements:

$$\begin{aligned} \mathbf{x}^{dd}(i) &= \lceil \frac{1}{\mathbf{R}(i,i)} (\mathbf{y}_{sic}(i) - \sum_{j=i+1}^{MN} \mathbf{R}(i,i) \mathbf{x}^{dd}(j)) \rfloor_o, \\ & i = MN-1, \cdots, 2, 1 \end{aligned} \tag{30}$$

where $\lceil . \rfloor_o$ denotes the nearest constellation points in $\mathcal{A}$.

B. Sorted QRD-based detectors

Since the lower-right positions of $\mathbf{R}$ are detected at first, they will influence the succeeding detection of the remaining elements. To minimize the error propagation, the reliability of these positions should be larger for the rest, i.e. the absolute value of the diagonal elements of $\mathbf{R}$ should be increasing. This is done by the QRD with column pivoting.

$$\mathbf{H}\mathbf{P} = \mathbf{Q}\mathbf{R}, \tag{31}$$

where $\mathbf{P}$ is a permutation matrix, i.e. $\mathbf{P}^{-1} = \mathbf{P}^{\mathbf{T}}$. By finding $\mathbf{P}$ in Eq.(31) such that the absolute values of the main diagonal elements of $\mathbf{R}$ are increasing, the detection process can be denoted as:

$$\mathbf{y}_{sic} = \mathbf{Q}^H \mathbf{y}^{dd} = \mathbf{R}(\mathbf{P}^{\mathbf{T}} \mathbf{x}^{dd}) + \tilde{\mathbf{n}}. \tag{32}$$

$\mathbf{x}_p^{dd} = \mathbf{P}^{\mathbf{T}} \mathbf{x}^{dd}$ is firstly detected by the SIC detector and $\mathbf{x}^{dd}$ is achieved by permutating $\mathbf{x}_p^{dd}$ according to $\mathbf{P}$.

The global optimal SQRD can be implemented by the exhausted pivoting of the columns, which requires $\frac{MN(MN-1)}{2}$ times of QRD. Its complexity forbids the application to a practical system. In this paper, a conventional reduced complexity with marginal performance loss is used[36], as illustrated in Algorithm (1).

Algorithm 1 SQRD with column pivoting

Input: $\mathbf{H} \in \mathbb{C}^{K \times MN}$
Output: $\mathbf{P} \in \mathbb{C}^{MN \times MN}, \mathbf{Q} \in \mathbb{C}^{K \times MN}, \mathbf{R} \in \mathbb{C}^{MN \times MN}$.
$\mathbf{R} = \mathbf{0}$, $\mathbf{Q} = \mathbf{H}$, $p = [1, 2, \cdots, MN]$.
for i = 1 to MN **do**
  $\omega_i = \mathbf{Q_i^T Q_i}$
end for
for i=1:MN **do**
  $k = argmin_{l=i:MN} \omega_l$
  exchange i-th and k-th column of $\mathbf{R}, \mathbf{Q}, \mathbf{p}$.
  $R_{i,i} = \sqrt{\omega_i}$
  $\mathbf{Q}_i = \mathbf{Q}_i / R_{i,i}$.
    for n = i+1:MN **do**
      $R_{i,n} = \mathbf{Q_i^H Q_i}$
      $\mathbf{Q}_n = \mathbf{Q}_n - R_{i,n}\mathbf{Q}_i$
      $\omega_n = \omega_n - |R_{i,n}|^2$
    end for
end for
Construct $\mathbf{P}$ according p.

C. Complexity analysis

The complexity of the QRD-based detection mainly comes from the QRD of the large size channel matrix $\mathbf{H}$, so here the QRD is analyzed. The complexity reduction is mainly two folds: 1) The number of items that need to be eliminated below the main diagonal of $\mathbf{H}$. 2) The sparsity of $\mathbf{H}$, which reduces the number of multiplications.

Since $\mathbf{H} = (\mathbf{I}_M \otimes \mathbf{F}_N)\mathbf{H}_o(\mathbf{I}_M \otimes \mathbf{F}_N^H)$, and $\mathbf{I}_M \otimes \mathbf{F}_N, \mathbf{I}_M \otimes \mathbf{F}_N^H$ are block diagonal matrix with block size $N \times N$, the overall non-zeros elements are determined by $\mathbf{H}_o$. According to Eq.24,

it can be shown that if $\tau_{max} \leq N$, then there are at most τ_{max} non-zeros items per column of $\mathbf{H}_o$, and since it is a Heisenberg matrix, there are at most $2N$ non-zero rows per N successive columns for $\mathbf{H}$, which form a block. Similarly, if $\tau_{max} \geq N$, there are at most $3N$ non-zero rows per N successive columns for $\mathbf{H}$, which form a block. In summary, the number of non-zero elements below the main diagonal is the trapezoid:

$$C_{nnz} = \begin{cases} \frac{(2N-1+N-1)N(M-1)}{2} + \frac{(N-1)N}{2} & \tau_{max} \leq N \\ \frac{(3N-1+2N-1)NM}{2} - 3N^2 & \tau_{max} > N \end{cases} \tag{33}$$

$C_{nnz} = \mathcal{O}(N^2M)$. However for the non-interleaved channel $\mathbf{H}$ of size $MN \times MN$, although $\mathbf{H}$ is also sparse, if there is one preceeding non-zero element of one row, the whole succeeding elements would be non-zeros after eliminating the first non-zeros elements. As a result, for the worst case the number of items need to be computed below the main diagonal is $\tilde{C}_{nnz} = \frac{(MN-1)MN}{2} - = \mathcal{O}(N^2M^2)$. Typically, M is much larger than N, so the complexity reduction is tremendous.

For the MMSE-SIC, it is noted that due to $\alpha\mathbf{I}_{MN}$ in the extended channel matrix $\tilde{\mathbf{H}}$, there is no complexity reduction based on the non-zero elements below the main diagonal, i.e. the number of rotation of Givens or Householder transformation required. However, the interleaved channel can still benefit the computation due to the sparse structure of $\mathbf{H}$. There are fixed $2N^2M$ and $3N^2M$ elements to be eliminated for $\tau_{max} \leq N$ and $\tau_{max} > N$ respectively for the interleaved version, while the non-interleaved version is M^2N^2. Thus, only $\mathcal{O}(\frac{1}{M})$ multiplications compared to the non-interleaved version.

V. Simulation

In this section, the simulation settings are presented, and the BER of different detection methods with different transmission configuration are performed.

A. Simulation configurations

The simulation configurations are illustrated in Table I. Two frame structures are evaluated in the simulation, where the number of Doppler and delay bins are set to $(8, 8)$ and $(8, 16)$ respectively. After the processing of ISFFT and IFFT, the CP is appended to the samples, which is larger than the maximum delay spread. It is noted that in our methods, at the transmitter

TABLE I: Simulation Parameters

Parameters	**Value**
Number of Blocks (N, M)	$(8, 8), (8, 16)$
Number of Frames	10000
Number of Paths (P)	3
Path Gain Model	Rayleigh
Power Normalized	Yes
Doppler Taps (v)	$[0, 2, 3]$
Delay Taps (τ)	$[0, 2, 3]$
Channel Estimation	Perfect
Modulation	4QAM, 16QAM, 64QAM
Modulation Mapping	Gray
Bits per Frame	$(128, 256, 384)$ for $(8, 8)$, $(256, 512, 768)$ for $(8, 16)$
SNR Range	$10 : 2 : 20$ for 4QAM,16QAM, $20 : 2 : 30$ for 64QAM
Detecting Methods	Message Passing (MP), ZF, ZF-SIC, MMSE-SIC, ZF-SQRD-SIC, MMSE-SQRD-SIC
α for MMSE	0.05

the CP is inserted after the temporal interleaving, and the CP is removed before the temporal deinterleaving at the receiver.

The LTV channel model considered in our paper consists of 3 paths, where the gain of each path is rayleigh, and both the delay and doppler taps are set to $[0, 2, 3]$. As a result, the theoretic maximum diversity is 3[37]. In order to fairly compare the performance of the detecting methods, perfect channel estimation is used for all the candidate methods in the paper.

Three modulation orders from low to high are simulated in the paper, i.e. 4QAM, 16QAM, 64QAM. All the constellations follow the gray mapping with power normalization. The SNR ranges of 4QAM and 16QAM are set from 10 to 20, and 64QAM are set from 20 to 30, with the step size being 2.

Six detecting methods are performed in the simulation, i.e. 1) Message passing (MP) detection in [38]. The code of MP is downloaded from the authors' website as-it-is except to align the configuration of channel model, resource block, and modulation orders. It is regarded as the near-optimal detecting method for OTFS[39]. 2) Zero-forcing (ZF), which is performed by

left multiplication the pseudo-inverse of the vectorized channel matrix. 3) ZF-SIC, which is performed by QRD of the vectorized channel matrix and the successive interference cancellation. 4) MMSE-SIC. Compared to ZF-SIC, the vectorized channel matrix is augmented with an additional diagonal block with $MN \times MN$. The parameter α of MMSE-SIC is set to 0.05 after several trials. 5) ZF-SQRD-SIC, which is the same as ZF-SIC except the sorting of the $\mathbf{R}_{ii}$ with the column pivoting. 6) MMSE-SQRD-SIC, which is also the same as MMSE-SIC except the sorting of $\mathbf{R}_{ii}$ with the column pivoting. The detecting orders of ZF-SQRD-SIC and MMSE-SQRD-SIC are determined by $\mathbf{P}$ according to Algorithm 1.

For each detecting method, 10^4 frames are generated, no channel coding is used. For every frame, $(128, 256, 386)$ and $(256, 386, 768)$ bits are uniformly generated for (4QAM,16QAM,64QAM) of (N,M)=(8,8), (8,16) respectively. The simulation environment is MATLAB with the parallel function for each SNR value to speed up the simulation. The whole simulation time for all methods except MP can be accomplished in less than 10 minutes with a 10-core Intel i9 CPU and 64GB RAM, and the MP method will be accomplished in about 3-4 hours for 64QAM.

B. Performance evaluation

The BER performance of different detection methods for 4QAM with block size (8,8) is illustrated in Figure 2. It is shown that the MP method performs the best in BER across all the SNRs, while ZF performs the worst. Besides, according to the gradient of the curves which represent the diversity of the detection methods achieved, it is shown that the ZF achieves the least diversity, and the MP achieves the most. With the SIC introduced, the diversity increase as in ZF-SIC and MMSE-SIC. MMSE-SIC and ZF-SIC achieve similar diversity while MMSE-SIC achieves better BER performance. Furthermore, with the sorted main diagonal elements of QRD, both the diversity and BER performance can be improved. It is shown that the QRD-based SIC detector can improve about 4dB gain at 10^{-2} compared to the non-SIC ZF detecting, while the sorted QRD can introduce an additional 2dB at 10^{-3} than the non-sorted detecting, i.e. ZF-SQRD-SIC vs. ZF-SIC.

The BER performance of different methods for 4QAM with block size (8,16) is illustrated in Figure 3. In this case, the MP method still performs the best in BER and diversity, and the sorted and non-sorted QRD based detectors can still bring benefits in both BER and diversity. It is shown that the diversity of sorted QRD-based detectors is quite close to the near-optimal MP method. Besides, compared to Figure 2 and 3, it is shown that with the increase of frame

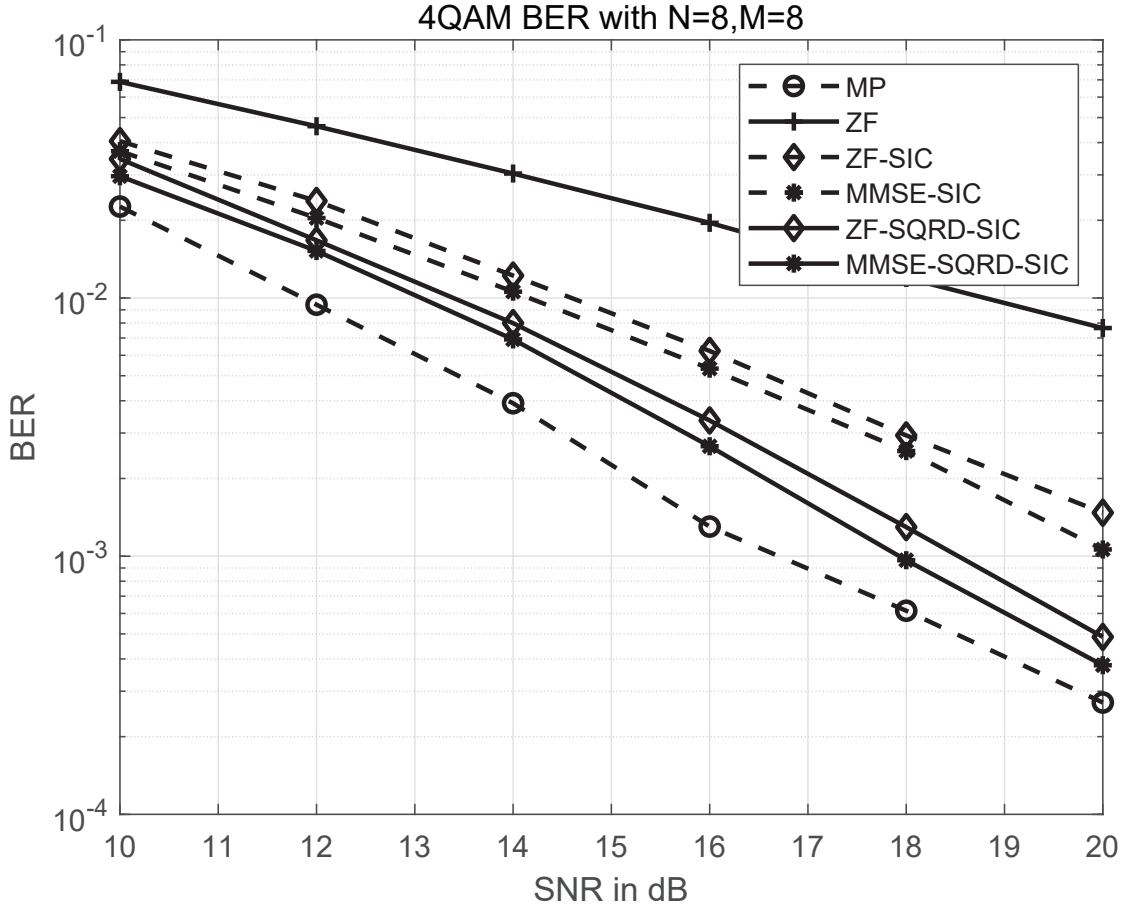

Fig. 2: Performance comparison of 4QAM with N=8, M=8

size, the performance of MP becomes better. This can be understood from the MP-decoding of LDPC-like codes, that the performance becomes better with the code length increasing.

The BER performance of different methods for 16QAM with block size (8,8) is illustrated in Figure 4. With the modulation order increasing, it is shown that the BER performance of the MP method decreases, which is even worst than the ZF method at low SNRs. While the sorted QRD methods, i.e. MMSE-SQRD-SIC and ZF-SQRD-SIC perform better in BER than the MP method. However, in this case, the diversity of the MP method is still the largest, and with the SNR increasing, the BER of MP drops very fast. The diversity of the sorted and non-sorted QRD-based methods performs similar to the 4QAM case, and the SIC and sorting procedures improve the achievable diversity. And the sorted version is very close to the MP method.

The BER performance of different methods for 16QAM with block size (8,16) is illustrated in Figure 5. Compared with Figure 5, it is shown that the BER of MP method becomes better with

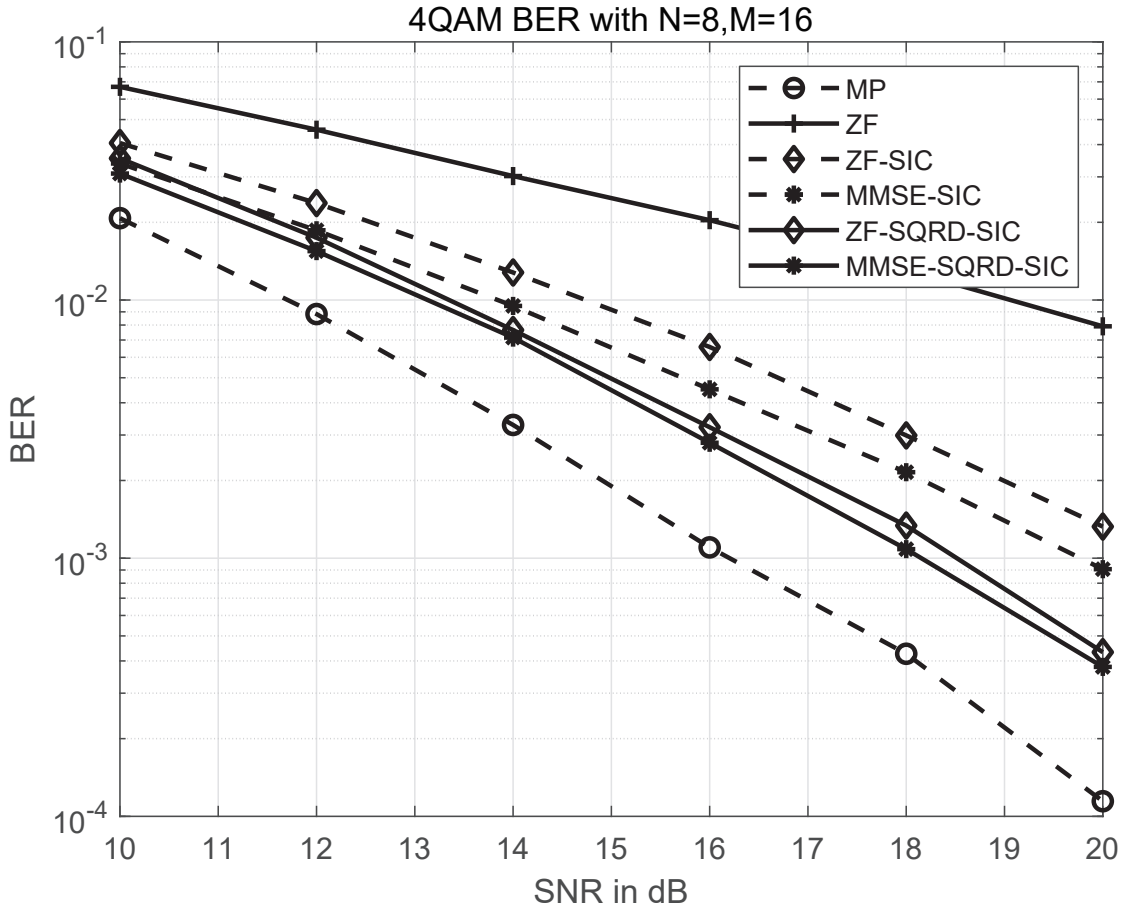

Fig. 3: Performance comparison of 4QAM with N=8, M=16

a larger frame size, while the BER performance of the other detecting methods is insensitive to the frame size.

The BER performance of different methods for 64QAM with block size (8,8) is illustrated in Figure 6. It is shown that with the modulation order increasing, the error floor appears for the MP method. Both the BER and diversity performance of the MP method does not increase with the SNR increasing. With higher modulation orders, the bits become error-prone, and the short girth in the factor graph degrade the demixture performance of MP method. The short girth would lead to the local-optimal of the consensus nodes, which severely degrades the detecting performance. While the QRD-based SIC methods, both the sorted and non-sorted versions, still perform well in the high modulation order. The improved performance of SIC is about 2dB at 10^{-2} than the non-SIC method, and the sorted versions introduce an additional 2dB.

The BER performance of different methods for 64QAM with block size (8,16) is illustrated

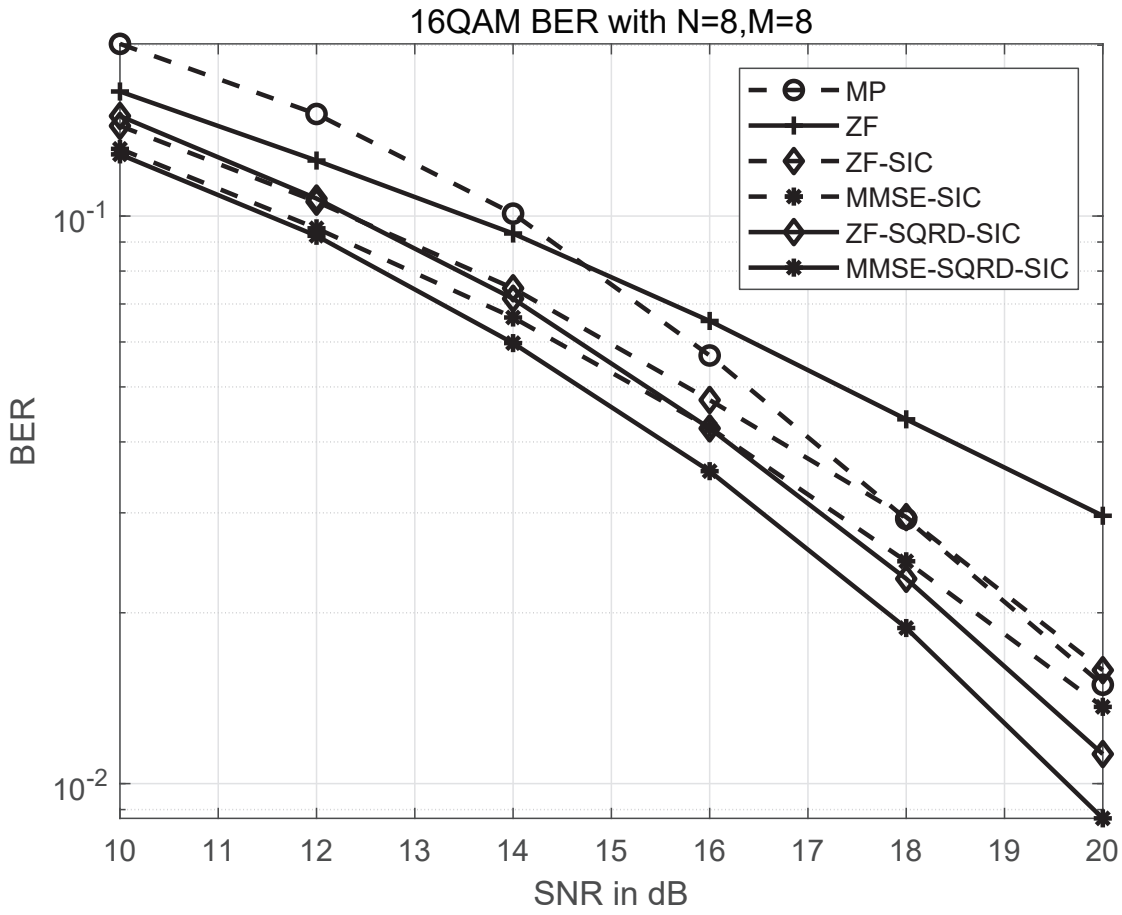

Fig. 4: Performance comparison of 16QAM with N=8, M=8

in Figure 7. It is shown that increasing the frame size would relieve the error floor of the MP method, however, it still performs the worst both in BER and diversity in this case. While the QRD-based methods perform similarly as in Figure 6.

According to the different performance of the MP method from low (4QAM), median (16QAM), and high (64QAM) modulation orders, it is shown that the special structure of the delay-doppler domain channel (i.e. short girth) will severely degrade the performance of MP method from median to high modulation orders. As a comparison, the QRD-based detecting methods, both the sorted and non-sorted versions, perform well across all the configurations. With the introduction of the temporal interleaving and de-interleaving, the QRD process can be computed fast and speed up, which is quite meaningful for the OTFS detection.

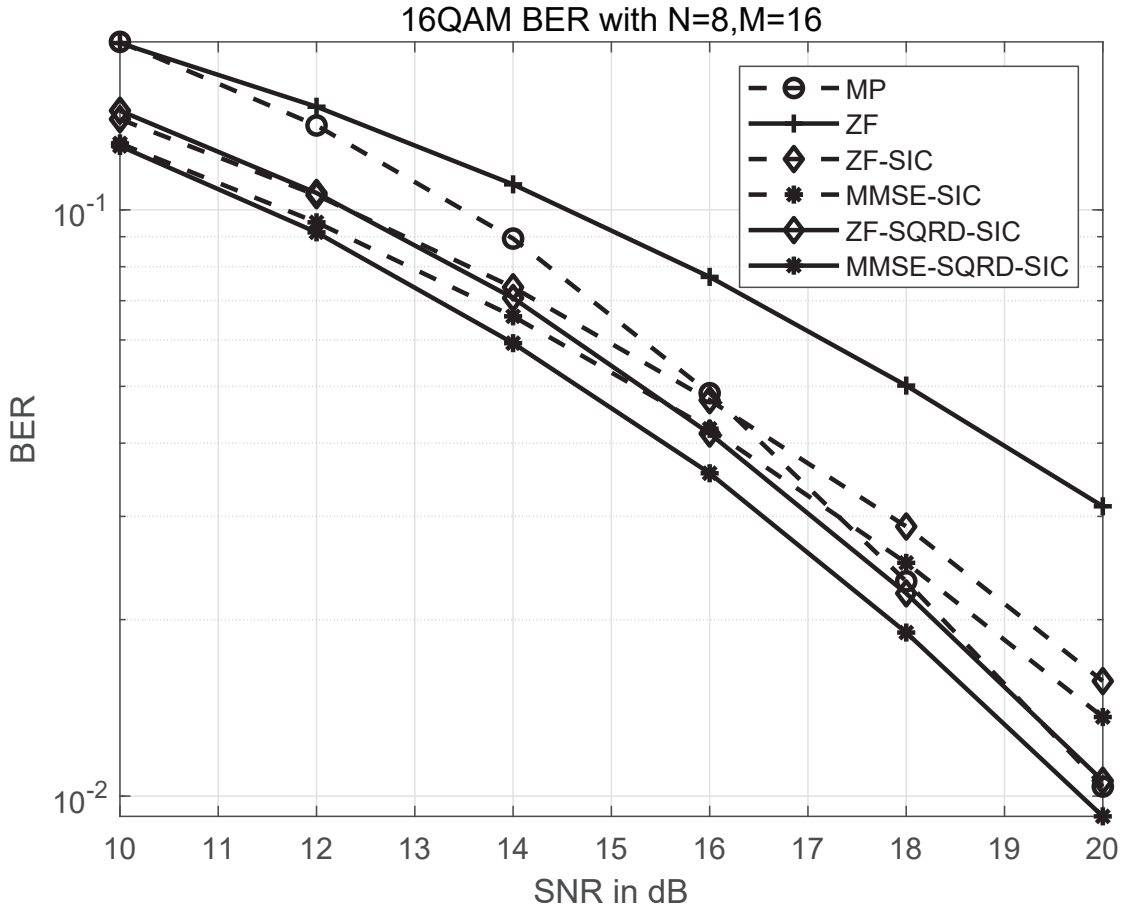

Fig. 5: Performance comparison of 16QAM with N=8, M=16

VI. Conclusion

This paper analyzed the delay-Doppler channel matrix structure and revealed the existence of short girth in the Tanner graph based on the sparse channel, which would severely degrade the performance of the MP method. By introducing the interleaver and deinterleaver at the transmitter and receiver, the channel matrix is in the form of a sparse upper block Heisenberg matrix, which can be QRD with fast algorithms. QRD-based SIC detectors were constructed in this paper. The simulation showed that the proposed detectors have performance and complexity advantages over the MP methods.

References

[1] 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.

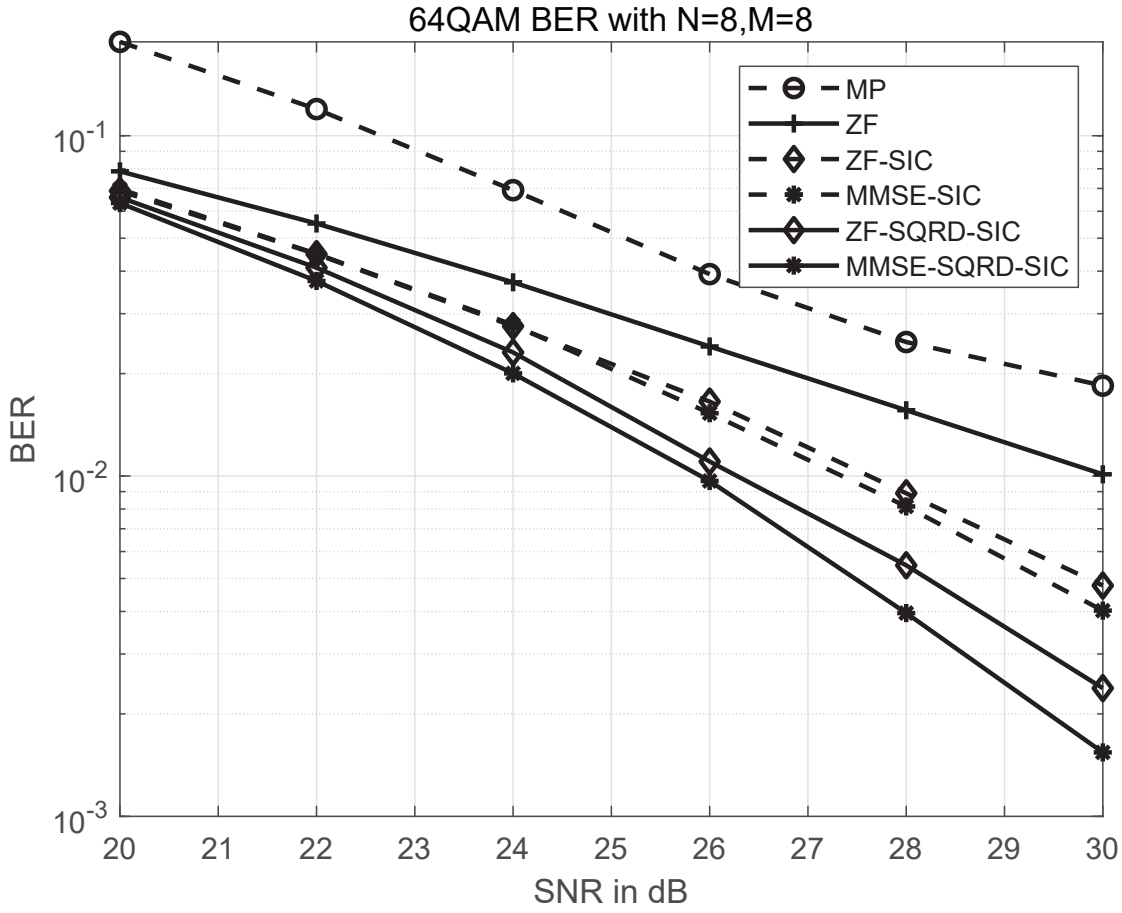

Fig. 6: Performance comparison of 64QAM with N=8, M=8

[2] K. R. Murali and A. Chockalingam, "On otfs modulation for high-doppler fading channels," in *2018 Information Theory and Applications Workshop (ITA)*, 2018, pp. 1–10.

[3] 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 Commun.*, vol. 28, no. 4, pp. 136–144, 2021.

[4] 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, 2018.

[5] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge University Press, 2008.

[6] 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 Commun. Lett.*, vol. 7, no. 3, pp. 344–347, 2018.

[7] G. D. Surabhi, R. M. Augustine, and A. Chockalingam, "Peak-to-average power ratio of otfs modulation," *IEEE Commun. Lett.*, vol. 23, no. 6, pp. 999–1002, 2019.

[8] X. Liu, X. Zhang, L. Zhang, P. Xiao, J. Wei, H. Zhang, and V. C. M. Leung, "Papr reduction using iterative clipping/filtering and admm approaches for ofdm-based mixed-numerology systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2586–2600, 2020.

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

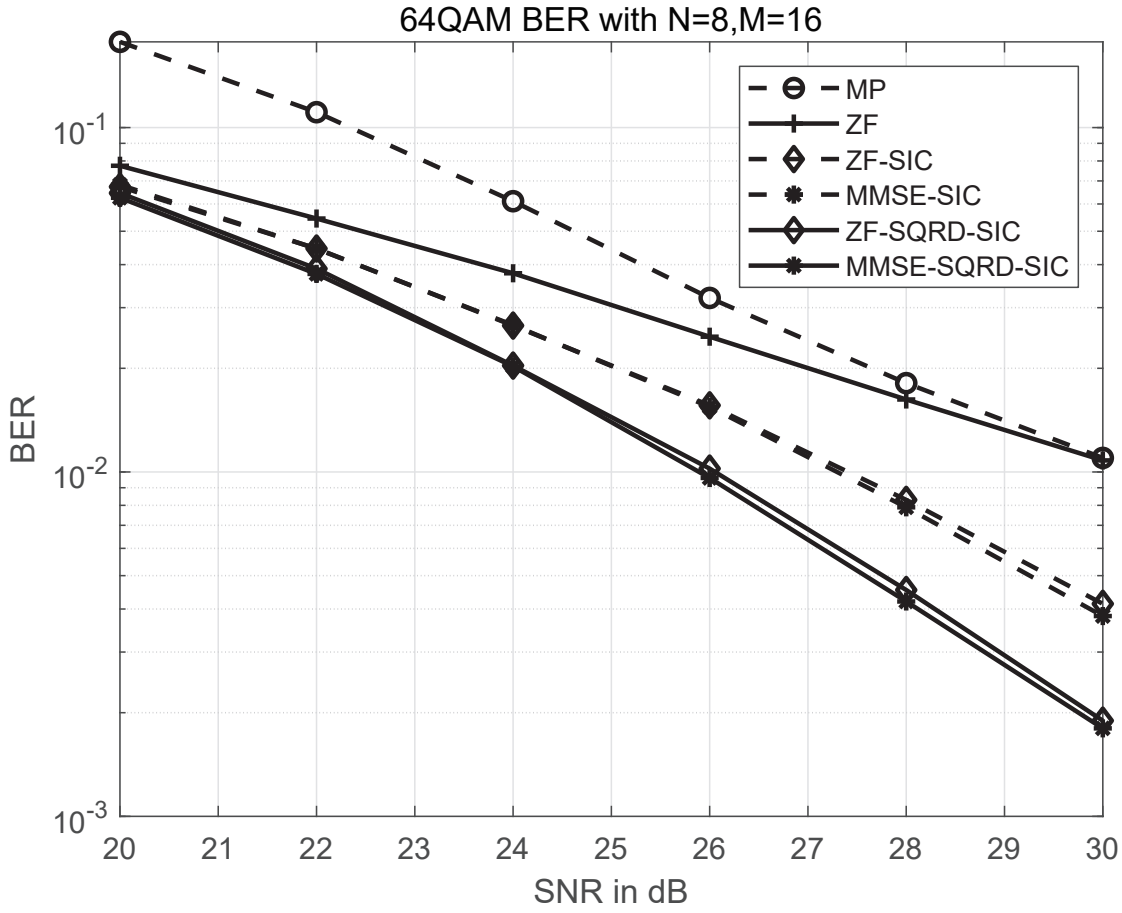

Fig. 7: Performance comparison of 64QAM with N=8, M=16

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

[11] L. Gaudio and G. Colavolpe, "Otfs vs. ofdm in the presence of sparsity: A fair comparison," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2021.

[12] R. Bomfin, A. Nimr, M. Chafii, and G. Fettweis, "A robust and low-complexity walsh-hadamard modulation for doubly-dispersive channels," *IEEE Commun. Lett.*, vol. 25, no. 3, pp. 897–901, 2021.

[13] P. Raviteja, K. T. Phan, and Y. Hong, "Embedded pilot-aided channel estimation for otfs in delay–doppler channels," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4906–4917, 2019.

[14] W. Shen, L. Dai, J. An, P. Fan, and R. W. Heath, "Channel estimation for orthogonal time frequency space (otfs) massive mimo," *IEEE Trans. Signal Process.*, vol. 67, no. 16, pp. 4204–4217, 2019.

[15] Y. Liu, S. Zhang, F. Gao, J. Ma, and X. Wang, "Uplink-aided high mobility downlink channel estimation over massive mimo-otfs system," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 1994–2009, 2020.

[16] H. Qu, G. Liu, L. Zhang, S. Wen, and M. A. Imran, "Low-complexity symbol detection and interference cancellation for otfs system," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1524–1537, 2021.

[17] S. Srivastava, R. K. Singh, A. K. Jagannatham, and L. Hanzo, "Bayesian learning aided simultaneous row and group sparse channel estimation in orthogonal time frequency space modulated mimo systems," *IEEE Trans. Commun.*, pp. 1–1, 2021.

[18] Z. Wei, W. Yuan, S. Li, J. Yuan, and D. W. K. Ng, "Transmitter and receiver window designs for orthogonal time-frequency space modulation," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2207–2223, 2021.

[19] G. D. Surabhi, R. M. Augustine, and A. Chockalingam, "On the diversity of uncoded otfs modulation in doubly-dispersive channels," *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, pp. 3049–3063, 2019.

[20] S. Tiwari, S. S. Das, and V. Rangamgari, "Low complexity lmmse receiver for otfs," *IEEE Commun. Lett.*, vol. 23, no. 12, pp. 2205–2209, 2019.

[21] P. Raviteja, Y. Hong, E. Viterbo, and E. Biglieri, "Effective diversity of otfs modulation," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 249–253, 2020.

[22] G. D. Surabhi and A. Chockalingam, "Low-complexity linear equalization for otfs modulation," *IEEE Commun. Lett.*, vol. 24, no. 2, pp. 330–334, 2020.

[23] W. Yuan, Z. Wei, J. Yuan, and D. W. K. Ng, "A simple variational bayes detector for orthogonal time frequency space (otfs) modulation," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 7976–7980, 2020.

[24] T. Thaj and E. Viterbo, "Low complexity iterative rake decision feedback equalizer for zero-padded otfs systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15 606–15 622, 2020.

[25] C. Jin, Z. Bie, X. Lin, W. Xu, and H. Gao, "A simple two-stage equalizer for otfs with rectangular windows," *IEEE Commun. Lett.*, vol. 25, no. 4, pp. 1158–1162, 2021.

[26] Z. Yuan, F. Liu, W. Yuan, Q. Guo, Z. Wang, and J. Yuan, "Iterative detection for orthogonal time frequency space modulation with unitary approximate message passing," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2021.

[27] S. Li, W. Yuan, Z. Wei, and J. Yuan, "Cross domain iterative detection for orthogonal time frequency space modulation," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2021.

[28] Y. Ge, Q. Deng, P. C. Ching, and Z. Ding, "Receiver design for otfs with a fractionally spaced sampling approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4072–4086, 2021.

[29] X. Zhang, J. Wang, H. Zhang, L. Li, M. Pan, and Z. Han, "Data-driven transportation network company vehicle scheduling with users' location differential privacy preservation," *IEEE Trans. Mobile Comput.*, pp. 1–1, 2021.

[30] Z. Ding, R. Schober, P. Fan, and H. Vincent Poor, "Otfs-noma: An efficient approach for exploiting heterogenous user mobility profiles," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7950–7965, 2019.

[31] M. Li, S. Zhang, F. Gao, P. Fan, and O. A. Dobre, "A new path division multiple access for the massive mimo-otfs networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 903–918, 2021.

[32] W. Yuan, Z. Wei, S. Li, J. Yuan, and D. W. K. Ng, "Integrated sensing and communication-assisted orthogonal time frequency space transmission for vehicular networks," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1515–1528, 2021.

[33] M. Fossorier, "Quasicyclic low-density parity-check codes from circulant permutation matrices," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1788–1793, 2004.

[34] A. Nimr, M. Chafii, M. Matthe, and G. Fettweis, "Extended gfdm framework: Otfs and gfdm comparison," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6.

[35] M. H. Gutknecht and T. Schmelzer, "Updating the qr decomposition of block tridiagonal and block hessenberg matrices," *Appl. Numer. Math.*, vol. 58, no. 6, p. 871–883, jun 2008.

[36] D. Wübben, R. Böhnke, J. Rinas, V. Kühn, and K. Kammeyer, "Efficient algorithm for decoding layered space-time codes," *Electron. Lett.*, vol. 37, pp. 1348–1350(2), October 2001.

[37] P. Raviteja, Y. Hong, E. Viterbo, and E. Biglieri, "Effective diversity of otfs modulation," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 249–253, 2020.

[38] 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, 2018.

[39] J. K. Francis, R. Mary Augustine, and A. Chockalingam, "Diversity and papr enhancement in otfs using indexing," in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, 2021, pp. 1–6.