Performance Analysis and Approximate Message Passing Detection of Orthogonal Time Sequency Multiplexing Modulation

Zeping Sui, *Student Member, IEEE*, Shefeng Yan, *Senior Member, IEEE*, Hongming Zhang, *Member, IEEE*, Sumei Sun, *Fellow, IEEE*, Yonghong Zeng, *Fellow, IEEE*, Lie-Liang Yang, *Fellow, IEEE*, and Lajos Hanzo, *Life Fellow, IEEE*

Abstract—In orthogonal time sequency multiplexing (OTSM) modulation, the information symbols are conveyed in the delay-sequency domain upon exploiting the inverse Walsh Hadamard transform (IWHT). It has been shown that OTSM is capable of attaining a bit error ratio (BER) similar to that of orthogonal time-frequency space (OTFS) modulation at a lower complexity, since the saving of multiplication operations in the IWHT. Hence we provide its BER performance analysis and characterize its detection complexity. We commence by deriving its generalized input-output relationship and its unconditional pairwise error probability (UPEP). Then, its BER upper bound is derived in closed form under both ideal and imperfect channel estimation conditions, which is shown to be tight at moderate to high signal-to-noise ratios (SNRs). Moreover, a novel approximate message passing (AMP) aided OTSM detection framework is proposed. Specifically, to circumvent the high residual BER of the conventional AMP detector, we proposed a vector AMP-based expectation-maximization (VAMP-EM) detector for performing joint data detection and noise variance estimation. The variance auto-tuning algorithm based on the EM algorithm is designed for the VAMP-EM detector to further improve the convergence performance. The simulation results illustrate that the VAMP-EM detector is capable of striking an attractive BER vs. complexity trade-off than the state-of-the-art schemes as well as providing a better convergence. Finally, we propose AMP and VAMP-EM turbo receivers for low-density parity-check (LDPC)-coded OTSM systems. It is demonstrated that our proposed VAMP-EM turbo receiver is capable of providing both BER and convergence performance improvements over the conventional AMP solution.

Zeping Sui and Shefeng Yan are with the Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China; and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: suizeping@mail.ioa.ac.cn; sfyan@mail.ioa.ac.cn).

Hongming Zhang is with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (email: zhanghm@bupt.edu.cn).

Sumei Sun and Yonghong Zeng are with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 138632 (e-mail: sunsm@i2r.a-star.edu.sg; yhzeng@i2r.a-star.edu.sg).

Lie-Liang Yang and Lajos Hanzo are with the Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: lly@ecs.soton.ac.uk; lh@ecs.soton.ac.uk).

Index Terms—Orthogonal time-frequency space (OTFS), orthogonal time sequency multiplexing (OTSM), approximate message passing (AMP), expectation-maximization (EM), turbo receiver.

I. INTRODUCTION

Next-generation wireless communication systems are expected to communicate reliably via high-mobility channels even in the face of severe inter-carrier interference (ICI) and inter-symbol interference (ISI) [1], [2]. With this in mind, orthogonal time-frequency space (OTFS) modulation has been proposed as an attractive candidate for next-generation wireless communications [3]–[6], where the information bits are modulated in the delay-Doppler (DD) domain. Then the two-dimensional orthogonal basis functions of inverse fast Fourier transform (IFFT) and FFT are employed along in the delay and Doppler domains, respectively, for spreading the DD domain symbols across the whole time-frequency (TF) domain. The most appealing aspect of OTFS is that it leverages the DD domain representation of the doubly-selective channel, where the dimension of the channel is reduced to the number of channel coefficients, implying that OTFS exhibits the sparsest channel representation [7], [8]. Moreover, the double-selective time-frequency (TF) channel is converted into the quasi-static DD domain channel based on the two-dimensional orthogonal basis functions, making OTFS capable of handling doubly-selective channels [9], [10]. However, the complexity of the OTFS system escalates, since the DD domain symbols are followed by the cascaded inverse symplectic finite Fourier transform (ISFFT)/SFFT and Heisenberg/Wigner transform pairs, especially for high numbers of time slots and subcarriers [11], [12].

As a relative of OTFS, orthogonal time sequency multiplexing (OTSM) modulation was first proposed in [13], [14], where sequency is defined as the number of the signal sign changes per second. In contrast to OTFS, the information symbols are arranged in the delay-sequency (DS) domain, and the delay-time (DT) domain frame is then obtained after invoking the inverse Walsh-Hadamard transform (IWHT) in the sequency domain in OTSM systems. The input-output relationship of OTSM proposed in [14] is derived based on the Fourier relationship between a pair of functions within the Bello system function family, i.e. the delay-spread function

TABLE I
CONTRASTING OUR CONTRIBUTIONS TO THE STATE-OF-THE-ART

Contributions	Our paper	[5]	[7]	[10]	[11]	[13]	[14]
OTSM	✓					✓	✓
BER performance analysis of OTSM	✓						
VAMP-based detection	✓						
Statistics of transmitted symbols	✓	✓	✓		✓		
EM noise variance estimation	✓						
Auto-tuning	✓						
Input-output relationship of generalized OTSM	✓						
Coded system	✓			✓	✓		✓
Turbo receiver	✓				✓		✓
Channel estimation imperfection	✓						✓

and the delay-Doppler-spread function [1]. Therefore, OTSM is also capable of exploiting the sparse representation of the time-varying multipath channel in the delay-sequency domain in a fashion reminiscent of the delay-Doppler domain channel representation of OTFS. Furthermore, OTSM may also be viewed as a single-carrier modulation scheme, where the ISI introduced by the delay spread and Doppler spread can be processed separately in delay and sequency domains, respectively [14]. Additionally, zero-padding (ZP) is integrated into OTSM systems to alleviate inter-block interference (IBI). As a benefit, OTSM is capable of attaining a similar bit error ratio (BER) performance to OTFS at a lower complexity in quasi-static and high-mobility channels, since the IWHT only needs addition and subtraction operations, rather than full-known multiplications.

Nonetheless, there are several open problems in the context of OTSM that require further investigation. Firstly, although the system model and the BER performances of several detectors designed for ZP-OTSM have been illustrated in [14], the generalized input-output relationship of OTSM dispensing with ZP and its formal BER analysis have not been unveiled. Furthermore, since the information symbols are mapped to the two-dimensional DS-domain grids, it is challenging to design low-complexity detectors having good performance. In [13], both a single-tap time domain (TD) and an iterative detector based on the Gauss-Seidel (GS) method have been proposed. However, the BER performance of the single tap detector degrades in high-mobility channels, while the GS detector includes matrix inversion at the cost of high complexity. It should be noted that none of the above detection algorithms exploits the sparsity of the DS-domain channel matrix, and neither do they exploit the statistics of the modulation schemes. We will fill this knowledge-gap and design efficient detectors based on approximate message passing (AMP) algorithms. Let us now review some of the existing works below to gain insight into designing AMP-based OTSM detectors.

The AMP algorithm that relies on loopy belief propagation (BP) relying on beneficial approximations has been proposed as a low-complexity signal recovery algorithm in [15]. The so-called "Onsager term" in the expressions of the AMP algorithm leads to a better sparsity-undersampling tradeoff than that of the iterative thresholding algorithms [16]. Nonetheless, the AMP algorithm was initially conceived for zero-mean independent and identically distributed (i.i.d.) Gaussian measurement matrices. For a general sensing matrix, the

performance of the AMP algorithm would suffer significantly. In [17], Rangan *et al.* proposed the vector AMP (VAMP) algorithm based on non-loopy factor graphs, which is suitable for right-orthogonally invariant matrices. Therefore, VAMP can handle a much broader class of matrices and achieves a significantly better mean square error (MSE) performance at a similar complexity to AMP. Given these compelling properties, the VAMP algorithm has been invoked for applications such as learning unknown hyper-parameters [18] and bilinear recovery [19]. However, the above-mentioned design philosophies have not been applied to the OTSM scheme.

Our novel contributions are boldly and explicitly contrasted to the existing literature in Table I at a glance, which are further detailed as follows:

- We first conceive the input-output relationship of generalized OTSM systems dispensing with ZP, which are different from the systems in [13], [14]. Then, the conditional pairwise error probability (PEP) of OTSM systems for a given DS-domain input-output relationship is derived by investigating the modified Euclidean distance of pairwise error events, and the unconditional PEP (UPEP) is further conceived based on PEP. Moreover, based on the UPEP and the union bounding technique, we derive a tight BER upper bound of OTSM systems in closed form, where the channel estimation imperfection is also considered.

- We propose a low-complexity VAMP-aided OTSM detector based on the *a priori* information of the modulation scheme, namely the VAMP-based expectation-maximization (VAMP-EM) detector. We first investigate the conventional AMP-aided detector. Then, based on the statistics of the transmitted symbols, the VAMP-aided detector is proposed by relying on the vector-valued factor graph. Thirdly, we exploit the EM algorithm to learn the noise variance to be incorporated in the VAMP detector. Hence, the proposed VAMP-EM detector provides more precise symbol vector estimates under unknown noise variance scenarios than the conventional AMP detector. As a further advance, an auto-tuning strategy is conceived for estimating the parameters γ_1 and γ_2 at every iteration in order to improve the efficiency of VAMP-EM. Simulation results are provided for characterizing the overall performance of the VAMP-EM detector. It is demonstrated that the proposed VAMP-EM achieves a significant BER and convergence rate improvement. Furthermore, the VAMP-EM detector is

capable of striking an attractive BER vs. complexity trade-off in uncoded OTSM systems.

- We extend the AMP and VAMP-EM detectors to turbo receivers designed for the low-density parity-check (LDPC) coded OTSM systems. Specifically, the attainable performances of the AMP and VAMP-EM turbo detectors conceived are evaluated at different LDPC-coded rates and different numbers of iterations, and the powerful semi-analytical tools of extrinsic information transfer (EXIT) charts are invoked to visually characterize the soft-information flow between the decoders and our proposed detectors. Our simulation results demonstrate that the proposed VAMP-EM turbo receiver attains both a considerable BER performance gain and a convergence speed-up.

The rest of the paper is organized as follows. The system model is portrayed in Section II, followed by our error performance analysis in Section III. In Section IV, the proposed AMP-aided detection algorithms are discussed, while the AMP and VAMP-EM-based turbo receivers are introduced in Section V. Our simulation results are illustrated in Section VI. Finally, our conclusions are offered in Section VII.

Notation: We use the following notation throughout this paper. $\mathbb{B}$ denotes the bit set consisting of $\{0, 1\}$; $\mathbb{E}[\cdot]$, $\text{tr}\{\cdot\}$, $[\cdot]_N$ and $\Re\{\cdot\}$ represent the expectation, trace, module-N and real part operators respectively; Matrices and vectors are denoted by upper- and lower-case boldface letters, respectively; For a length-n vector $\boldsymbol{u}$, the empirical averaging operator can be given as $\langle \boldsymbol{u} \rangle = \frac{1}{N}\sum_{n=1}^{N} u(n)$; $\text{vec}(\boldsymbol{A})$ denotes the vector formulated by stacking the columns of $\boldsymbol{A}$ to obtain a single column vector matrix, and $\text{vec}^{-1}(\boldsymbol{a})$ denotes the inverse vectorization operation to form the original matrix; $\otimes$ is the Kronecker product of two matrices; $I_0(\cdot)$ the zero-order modified Bessel function of the first kind; $\mathcal{CN}(\boldsymbol{a}, \boldsymbol{B})$ denotes the complex Gaussian distribution with mean vector $\boldsymbol{a}$ and covariance matrix $\boldsymbol{B}$; For a matrix $\boldsymbol{A}$, $(\boldsymbol{A})^T$, $(\boldsymbol{A})^*$, $(\boldsymbol{A})^H$, $(\boldsymbol{A})^{-1}$ and $|\boldsymbol{A}|^2$ represent its transpose, conjugate, conjugate transpose, inverse and element-wise modulo squared operation respectively; $\boldsymbol{\mathcal{W}}_N$ stands for the N-point IWHT matrix; $\boldsymbol{I}_N$ and $\boldsymbol{e}_N(n)$ denote the N-dimensional identity matrix and its nth column; $\boldsymbol{0}_{u \times v}$ is the $u \times v$ zero matrix; $Q(x) = \int_x^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt$ and $\delta(\cdot)$ represent the Q-function and the delta function; $\boldsymbol{x} \cdot \boldsymbol{y}$ and $\boldsymbol{x} \cdot / \boldsymbol{y}$ denotes the element-wise product and division between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$; $\boldsymbol{1}$ and $\boldsymbol{0}$ denote the all-ones and all-zeros vectors with a proper length, respectively; $H(q) \triangleq -\int q(x)\ln q(x)\mathrm{d}x$ denotes the differential entropy of $x \sim q$; $D_{\text{KL}} \triangleq \int q(x)\ln\frac{q(x)}{p(x)}\mathrm{d}x$ is the *Kullback-Leibler divergence* from p to q; $\mathcal{U}[a,b]$ represents the uniform distribution in the interval $[a, b]$.

II. SYSTEM MODEL

A. Transmitter Description

We consider a ZP-OTSM system having a bandwidth $B = M\Delta f$ and the frame duration $T_f = NT$, where T represents the symbol duration and Δf is the sampling frequency interval, while M and N denote the maximum value of delay and sequency indices. Moreover, the system is critically sampled,

i.e., we have $T\Delta f = 1$. Throughout this section, we denote the number of data subcarriers by $L_D = M - L_{ZP}$, where L_{ZP} represents the length of ZP. The OTSM system is shown in Fig. 1. At the transmitter, an i.i.d information bit sequence $\boldsymbol{b} \in \mathbb{B}^{RL_b}$ is encoded by a rate-R LDPC encoder. Hence the LDPC-coded bit sequence $\boldsymbol{c} = [\boldsymbol{c}_0, \ldots, \boldsymbol{c}_{J-1}] \in \mathbb{B}^{L_b}$ is obtained after interleaving, where $\boldsymbol{c}_j = [c_j(1), \ldots, c_j(\log_2 K)]$ for $j = 0, \ldots, J-1$ and $L_b = NL_D\log_2 K = J\log_2 K$. Here K is the modulation order and the normalized K-ary alphabet is denoted $\mathcal{A} = \{a_1, \ldots, a_K\}$, with each constellation point a_k corresponds to a bit pattern $\boldsymbol{s}_k = [s_k(1), \ldots, s_k(\log_2 K)]$ for $k = 1, \ldots, K$. Then the coded bit sequence is mapped onto J quadrature amplitude modulation (QAM) symbols that can be formulated as $\boldsymbol{x}_D = [\boldsymbol{x}_0^T, \ldots, \boldsymbol{x}_{L_D-1}^T]^T \in \mathbb{C}^{NL_D \times 1}$, where $\boldsymbol{x}_{l_d} = [x_{l_d}(0), x_{l_d}(1), \ldots, x_{l_d}(N-1)]^T \in \mathbb{C}^{N \times 1}$ for $l_d = 0, \ldots, L_D - 1$, and we have $x_{l_d}(n) \in \mathcal{A}, \forall l_d, n$. The DS-domain data matrix $\boldsymbol{X}_{DS}$ can be obtained by arranging symbol vectors on the DS-domain lattice $\Gamma = \left\{\left(\frac{l}{M\Delta f}, \frac{n}{NT}\right), l = 0, \ldots, L_D - 1, n = 0, \ldots, N-1\right\}$ as $\boldsymbol{X}_{DS} = [\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{L_D-1}]^T \in \mathbb{C}^{L_D \times N}$. The relationship between the sampling grids in the DS-domain, the DT-domain, and the TD is demonstrated in Fig. 2, which is detailed below. Firstly, we obtain the transmitted DS-domain matrix after inserting the ZP, which can be written as $\boldsymbol{X} = \boldsymbol{\Upsilon}_{ZP}\boldsymbol{X}_{DS} \in \mathbb{C}^{M \times N}$, where $\boldsymbol{\Upsilon}_{ZP} = [\boldsymbol{I}_{L_D}, \boldsymbol{0}_{L_D \times L_{ZP}}]^T$ is the $(M \times L_D)$-element mapping matrix. Then IWHT is applied to every row of $\boldsymbol{X}$ to obtain the DT-domain transmitted matrix, yielding $\boldsymbol{X}_{DT} = \boldsymbol{X}\boldsymbol{\mathcal{W}}_N$. The elements of the $(N \times M)$-dimensional WHT matrix $\boldsymbol{\mathcal{W}}_N$ are given as $\mathcal{W}_N(n, m) = \tilde{W}(n, m/N + 0.5/N)/\sqrt{N}$, where $\tilde{W}(n, \lambda)$ represents the continuous Walsh functions within $0 \leq \lambda \leq 1$ [14]. Hence, the TD transmitted signal can be attained as $\boldsymbol{s} = \text{vec}(\boldsymbol{X}_{DT}) = \boldsymbol{P}(\boldsymbol{I}_M \otimes \boldsymbol{\mathcal{W}}_N)\boldsymbol{x}$, where $\boldsymbol{x} = [\boldsymbol{x}_D^T, \boldsymbol{0}_{1 \times L_{ZP}}]^T \in \mathbb{C}^{MN \times 1}$ and the $\boldsymbol{P}$ is the $(MN \times MN)$-dimensional row-column permutation matrix, which is also known as the perfect shuffle matrix [20], formulated as:

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{I}_N \otimes \boldsymbol{e}_M^T(0) \\ \boldsymbol{I}_N \otimes \boldsymbol{e}_M^T(1) \\ \vdots \\ \boldsymbol{I}_N \otimes \boldsymbol{e}_M^T(M-1) \end{bmatrix}. \tag{1}$$

Based on the fact that $\boldsymbol{P}^T = \boldsymbol{P}^{-1}$ and $\boldsymbol{A} \otimes \boldsymbol{B} = \boldsymbol{P}(\boldsymbol{B} \otimes \boldsymbol{A})\boldsymbol{P}^T$ for square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we have

$$\boldsymbol{s} = (\boldsymbol{\mathcal{W}}_N \otimes \boldsymbol{I}_M)(\boldsymbol{P}\boldsymbol{x}). \tag{2}$$

B. Channel Model

Overall a time-varying multipath channel with P taps is considered, whose channel impulse response (CIR) can be given in the delay-Doppler-spread Bello function form as [1] :

$$h(\tau, \nu) = \sum_{i=1}^{P} h_i \delta(\tau - \tau_i)\delta(\nu - \nu_i), \tag{3}$$

where h_i, τ_i and ν_i denote the complex-valued path gain, symbol duration-normalized delay and normalized Doppler shift corresponding to the ith path, respectively. Here, we
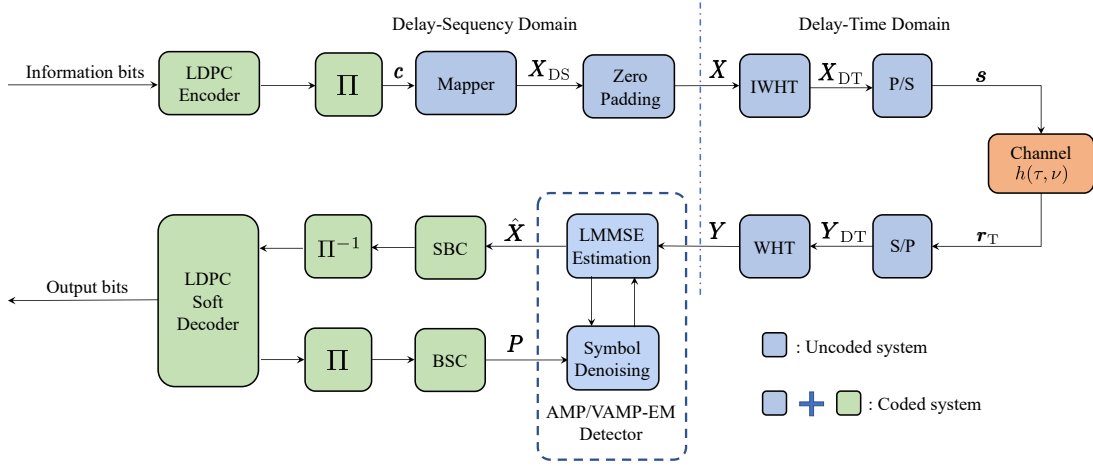
Fig. 1. Illustration of the uncoded and LDPC coded OTSM systems, where Π and Π^{-1} denote the turbo interleaver and deinterleaver, respectively.
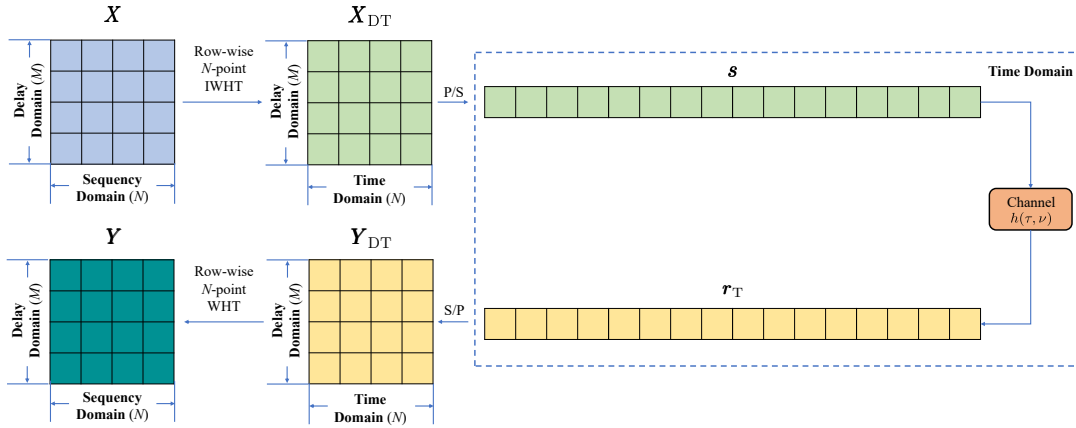

Fig. 2. Relationship between the grids in the DS-domain, DT-domain, and time domain of the OTSM systems with $M = 4$ and $N = 4$.

assume that $h_i \sim \mathcal{CN}(0, 1/P), \forall i$, which is independent of τ_i and ν_i [11]. The sampling values of delay shifts and Doppler shifts are denoted as $l_i = a_i + \alpha_i$ and $k_i = b_i + \beta_i$, respectively, where a_i and b_i represent the integer-valued delay and Doppler indices, while $\alpha_i, \beta_i \in \mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$ are the non-integer fractional components. Therefore, the normalized delay and Doppler shifts are defined as $\tau_i = \frac{l_i}{M\Delta f}$, $\nu_i = \frac{k_i}{NT}$. The maximum delay of the channel is $\tau_{max} = l_{max}/M\Delta f$ where $l_{max} = \max\{\mathcal{L}\} = \max\{l_1, \ldots, l_P\}$. Furthermore, we set $L_{ZP} = l_{max} + 1$ for mitigating the IBI between neighboring data blocks. Based on the intrinsic Fourier relationships among the Bello functions, the continuous time-varying CIR can be expressed with the aid of the input delay-spread function as [1]

$$h(\tau, t) = \int h(\tau, \nu)e^{j2\pi\nu(t-\tau)}d\nu = \sum_{i=1}^{P} h_i e^{j2\pi\nu_i(t-\tau_i)}, \quad (4)$$

which is processed by sampling at the grid points $\{\tau = \frac{l}{M\Delta f}, t = \frac{q}{M\Delta f}\}$, yielding

$$h(l, q) = \sum_{i=1}^{P} h_i z_i^{(q-l)}\delta(l - l_i), \quad (5)$$

where $z_i = e^{j2\pi\frac{k_i}{MN}}$ and $q = 0, 1, \ldots, MN - 1$.

C. Received Signals

Given the delay-Doppler-spread function $h(\tau, \nu)$, the received TD signal can be written as [1]

$$r_T(t) = \int\int h(\tau, \nu)s(t - \tau)e^{j2\pi\nu(t-\tau)}d\tau d\nu + n(t), \quad (6)$$

where $n(t)$ is the complex additive white Gaussian noise (AWGN) with zero mean and variance $N_0 = 1/\gamma_n$. Here, γ_n denotes the inverse noise variance, i.e., the noise precision. Therefore, by sampling the received signal at $\{t = \frac{q}{M\Delta f}, q = 0, 1, \ldots, MN - 1\}$, we can formulate the corresponding discrete-time received signals based on (5) and (6) as

$$r_T(q) = \sum_{l \in \mathcal{L}} h(l, q)s(q - l) + n(q), \quad (7)$$

where $s(q) = s(t)|_{t=\frac{q}{M\Delta f}}$ and $n(q) = n(t)|_{t=\frac{q}{M\Delta f}}$. Therefore, the (MN)-samples of the TD received signal and noise can be written as $\mathbf{r}_T = [r_T(0), \ldots, r_T(MN - 1)]^T$ and $\mathbf{n} = [n(0), \ldots, n(MN - 1)]^T$, respectively. Hence, based on the TD input-output relationship of (6), the corresponding received symbol vector is given by

$$\mathbf{r}_T = \mathbf{H}_T\mathbf{s} + \mathbf{n}, \quad (8)$$

where $\mathbf{H}_T = \text{diag}[\mathbf{H}_{T,0}, \ldots, \mathbf{H}_{T,N-1}] \in \mathbb{C}^{MN \times MN}$ is the effective TD channel matrix, and $\mathbf{H}_{T,n} \in \mathbb{C}^{M \times M}$ is the nth
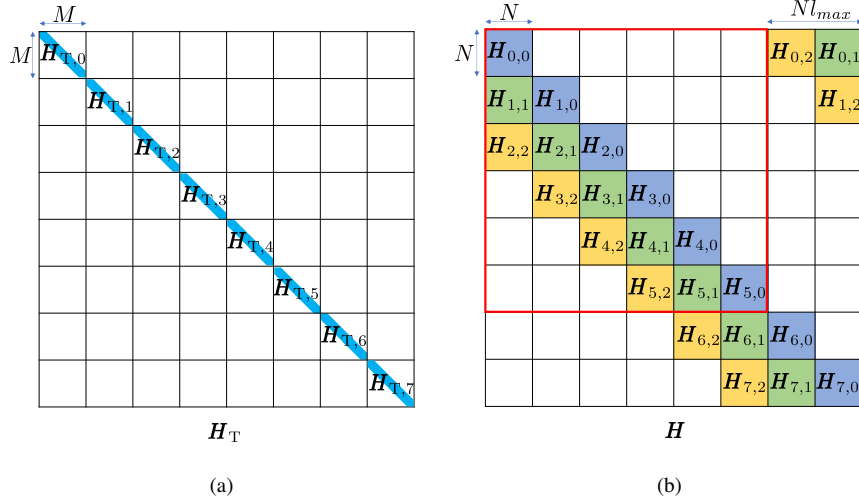
Fig. 3. Illustration of the effective time-domain channel matrix $\boldsymbol{H}_{\mathrm{T}}$ and delay-sequency domain channel matrix $\boldsymbol{H}$ with $M = N = 8$ and $l_{\max} = 2$.

block channel matrix for $n = 0, \ldots, N-1$. The non-zero valued elements of $\boldsymbol{H}_{\mathrm{T}}$ can be expressed as $H_{\mathrm{T}}(q, q-l) = h(l, q)$. By rewriting all the elements of $\boldsymbol{H}_{\mathrm{T}}$ as $H(a, b)$, it can be readily shown that $l_{\max} = \max\{a - b\}$, $\forall a, b$. Therefore, as shown in Fig. 3 (a), the bandwidth of $\boldsymbol{H}_{\mathrm{T}}$ is $l_{\max}$ and there are $L \leq l_{\max} + 1$ non-zero values in each row and column of $\boldsymbol{H}_{\mathrm{T}}$ [14]. After the inverse vectorization process, the received DT-domain symbol matrix can be formulated as $\boldsymbol{Y}_{\mathrm{DT}} = \mathrm{vec}^{-1}(\boldsymbol{r}_{\mathrm{T}}) = [\boldsymbol{y}_{\mathrm{DT},0}, \ldots, \boldsymbol{y}_{\mathrm{DT},M-1}]^T$. The received DS-domain matrix $\boldsymbol{Y}$ is obtained by applying the WHT to each row of $\boldsymbol{Y}_{\mathrm{DT}}$, yielding $\boldsymbol{Y} = \boldsymbol{Y}_{\mathrm{DT}}\boldsymbol{W}_N = [\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{M-1}]^T$. Hence, the received DS-domain symbol vector is given by

$$\boldsymbol{y} = (\boldsymbol{I}_M \otimes \boldsymbol{W}_N)(\boldsymbol{P}^T \boldsymbol{r}_{\mathrm{T}}), \tag{9}$$

where $\boldsymbol{y} = [\boldsymbol{y}_0^T, \ldots, \boldsymbol{y}_{M-1}^T]^T$. Finally, based on (2), (8) and (9), the DS-domain input-output relationship can be expressed as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \bar{\boldsymbol{n}}, \tag{10}$$

where the DS-domain channel matrix and the equivalent AWGN vector can be respectively formulated as $\boldsymbol{H} = (\boldsymbol{I}_M \otimes \boldsymbol{W}_N)(\boldsymbol{P}^T \boldsymbol{H}_{\mathrm{T}} \boldsymbol{P})(\boldsymbol{I}_M \otimes \boldsymbol{W}_N)$ and $\bar{\boldsymbol{n}} = (\boldsymbol{I}_M \otimes \boldsymbol{W}_N)(\boldsymbol{P}^T \boldsymbol{n})$. Since the transform matrix $(\boldsymbol{I}_M \otimes \boldsymbol{W}_N)\boldsymbol{P}^T$ is unitary, we can readily show that $\bar{n}(q) \sim \mathcal{CN}(0, \gamma_n^{-1})$, $\forall q$. Therefore, the average SNR per symbol can be expressed as $\gamma_s = \gamma_n$. As illustrated in Fig. 3 (b), the strictly upper triangular submatrices of $\boldsymbol{H}$ can be ignored since ZP is used, while the bandwidth of $\boldsymbol{H}$ is $Nl_{\max}$, and each row and column of $\boldsymbol{H}$ have NL non-zero elements. The sub-matrices in $\boldsymbol{H}$ can be written as $\boldsymbol{H}_{m,l} = \boldsymbol{W}_N \bar{\boldsymbol{H}}_{m,l} \boldsymbol{W}_N$ [14], where $\bar{\boldsymbol{H}}_{m,l} = \mathrm{diag}[\bar{h}_{m,l}(0), \ldots, \bar{h}_{m,l}(N-1)]$ and $\bar{h}_{m,l} = h(l, m+nM)$ for $m = 0, \ldots, M-1$ and $n = 0, \ldots, N-1$.

III. Error Performance Analysis

In this section, we investigate the BER performance of generalized OTSM systems dispensing with ZP, which differs from the ZP-OTSM systems in Section II. Throughout this section, we assume that perfect channel state information (CSI) is available at the receiver, and the optimum ML detector detailed in Section IV-A is employed. For notational consistency, we also use $\{\boldsymbol{s}, \boldsymbol{r}_{\mathrm{T}}, \boldsymbol{x}, \boldsymbol{y}\}$ to denote the TD and DS-domains transmitted and received symbol vectors, respectively. We commence by detailing the input-output relationship of the generalized OTSM systems. Then the conditional PEP (CPEP) and the UPEP of the OTSM system are derived. Finally, the asymptotically tight upper bound on the BER of generalized OTSM systems is derived.

To alleviate the ISI between OTSM frames, only a cyclic prefix (CP) length of $L_{\mathrm{CP}} = l_{\max} + 1$ is appended to the transmitted signal. The TD input-output relationship given by (7) can be further formulated as

$$r_{\mathrm{T}}(q) = \sum_{i=1}^{P} h_i e^{j 2\pi \frac{k_i(q - l_i)}{MN}} s([q - l_i]_{MN}) + n(q). \tag{11}$$

where $q = 0, 1, \ldots, MN - 1$. Therefore, (11) can be written in a vectorial form as

$$\boldsymbol{r}_{\mathrm{T}} = \boldsymbol{G}\boldsymbol{s} + \boldsymbol{n}, \tag{12}$$

Specifically, the TD channel matrix can be derived as $\boldsymbol{G} = \sum_{i=1}^{P} h_i \boldsymbol{\Pi}^{l_i} \boldsymbol{\Delta}^{k_i}$, where the permutation matrix $\boldsymbol{\Pi}$ can be obtained by employing a cyclic forward shift to the rows of $\boldsymbol{I}_{MN}$ [8], and $\boldsymbol{\Delta} = \mathrm{diag}[z^0, \ldots, z^{MN-1}]$ with $z = e^{j\frac{2\pi}{MN}}$. Given the transformation relationships shown in Fig. 2, by substituting (2) and (9) into (12), the DS-domain input-output relationship can be derived as

$$\boldsymbol{y} = \bar{\boldsymbol{H}}\boldsymbol{x} + \bar{\boldsymbol{n}}, \tag{13}$$

where $\boldsymbol{x} \in \mathcal{A}^{NM \times 1}$. Specifically, the DS-domain channel matrix can be written as

$$\bar{\boldsymbol{H}} = \sum_{i=1}^{P} h_i (\boldsymbol{I}_M \otimes \boldsymbol{W}_N)(\boldsymbol{P}^T \boldsymbol{\Pi}^{l_i} \boldsymbol{\Delta}^{k_i} \boldsymbol{P})(\boldsymbol{I}_M \otimes \boldsymbol{W}_N). \tag{14}$$

A. Ideal Channel Conditions

Substituting (14) into (13), it can be shown that (13) can be further formulated as

$$\boldsymbol{y} = \boldsymbol{\Phi}(\boldsymbol{x})\boldsymbol{h} + \bar{\boldsymbol{n}}, \tag{15}$$

where $\boldsymbol{h} = [h_1, \ldots, h_P]^T \in \mathbb{C}^{P \times 1}$ is the channel coefficient vector. The concatenated equivalent codeword matrix $\boldsymbol{\Phi}(\boldsymbol{x}) \in \mathbb{C}^{MN \times P}$ can be expressed as

$$\boldsymbol{\Phi}(\boldsymbol{x}) = [\underbrace{\boldsymbol{D}_1\boldsymbol{x}}_{MN \times 1} \quad \boldsymbol{D}_2\boldsymbol{x} \quad \ldots \quad \boldsymbol{D}_P\boldsymbol{x}], \qquad (16)$$

where $\boldsymbol{D}_p$ is given by $\boldsymbol{D}_p = (\boldsymbol{I}_M \otimes \boldsymbol{\mathcal{W}}_N)(\boldsymbol{P}^T\boldsymbol{\Pi}^{l_i}\boldsymbol{\Delta}^{k_i}\boldsymbol{P})(\boldsymbol{I}_M \otimes \boldsymbol{\mathcal{W}}_N)$, for $p = 1, \ldots, P$.

Let us consider the pairwise error event $\{\boldsymbol{x}^c \to \boldsymbol{x}^e\}$, where $\boldsymbol{x}^c$ is the transmitted symbol vector, and $\boldsymbol{x}^e$ is the error-infested received symbol vector. We define the error vector as $\boldsymbol{e} = \boldsymbol{x}^c - \boldsymbol{x}^e$. Hence, the modified Euclidean distance under this event can be expressed as

$$\delta = ||\boldsymbol{\Phi}(\boldsymbol{e})\boldsymbol{h}||_2^2 = \boldsymbol{h}^H\boldsymbol{\Phi}(\boldsymbol{e})^H\boldsymbol{\Phi}(\boldsymbol{e})\boldsymbol{h} = \boldsymbol{h}^H\boldsymbol{\Theta}\boldsymbol{h}, \quad (17)$$

where the codeword difference matrix can be found with the aid of (15) and (16), which is given by

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{e}^H\boldsymbol{D}_1^H\boldsymbol{D}_1\boldsymbol{e} & \cdots & \boldsymbol{e}^H\boldsymbol{D}_1^H\boldsymbol{D}_P\boldsymbol{e} \\ \vdots & \ddots & \vdots \\ \boldsymbol{e}^H\boldsymbol{D}_P^H\boldsymbol{D}_1\boldsymbol{e} & \cdots & \boldsymbol{e}^H\boldsymbol{D}_P^H\boldsymbol{D}_P\boldsymbol{e} \end{bmatrix}. \qquad (18)$$

The CPEP can be formulated as [21], [22]

$$\begin{aligned} P(\boldsymbol{x}^c \to \boldsymbol{x}^e|\boldsymbol{h}) &= P\left[||\boldsymbol{y} - \boldsymbol{\Phi}(\boldsymbol{x}^e)\boldsymbol{h}||_2^2 \leq ||\boldsymbol{y} - \boldsymbol{\Phi}(\boldsymbol{x}^c)\boldsymbol{h}||_2^2\right] \\ &= P[2\Re\{\boldsymbol{y}^H\boldsymbol{\Phi}(\boldsymbol{x}^e)\boldsymbol{h}\} - 2\Re\{\boldsymbol{y}^H\boldsymbol{\Phi}(\boldsymbol{x}^c)\boldsymbol{h}\} \\ &\geq ||\boldsymbol{\Phi}(\boldsymbol{x}^e)\boldsymbol{h}||_2^2 - ||\boldsymbol{\Phi}(\boldsymbol{x}^c)\boldsymbol{h}||_2^2]. \qquad (19) \end{aligned}$$

Upon exploiting (15) and substituting $\boldsymbol{\Phi}(\boldsymbol{e}) = \boldsymbol{\Phi}(\boldsymbol{x}^c) - \boldsymbol{\Phi}(\boldsymbol{x}^e)$ into (19), as well as following some rearrangement, (19) can be further rewritten as

$$P(\boldsymbol{x}^c \to \boldsymbol{x}^e|\boldsymbol{h}) = P\left[\Re\{\bar{\boldsymbol{n}}^H\boldsymbol{\Phi}(\boldsymbol{e})\boldsymbol{h}\} \geq \frac{||\boldsymbol{\Phi}(\boldsymbol{e})\boldsymbol{h}||_2^2}{2}\right]. \quad (20)$$

It can be shown that $\Re\{\bar{\boldsymbol{n}}^H\boldsymbol{\Phi}(\boldsymbol{e})\boldsymbol{h}\}$ obeys Gaussian distribution with zero mean and a variance of $\frac{||\boldsymbol{\Phi}(\boldsymbol{e})\boldsymbol{h}||_2^2}{2\gamma_s}$. Therefore, based on (17), the CPEP of (19) can be expressed as

$$P(\boldsymbol{x}^c \to \boldsymbol{x}^e|\boldsymbol{h}) = Q\left(\sqrt{\frac{\delta\gamma_s}{2}}\right) \leq \frac{1}{2}\exp\left(-\frac{\delta\gamma_s}{4}\right), \quad (21)$$

where the inequality is obtained based on the well-known Chernoff bound $Q(x) \leq \frac{1}{2}\exp(-\frac{x^2}{2})$ [23]. Explicitly, $\boldsymbol{\Theta}$ is a positive semidefinite Hermitian matrix. Let rank$(\boldsymbol{\Theta}) = r$, where $1 \leq r \leq P$. Then, when the eigenvalues are sorted in descending order and eigenvectors of $\boldsymbol{\Theta}$ are denoted as $\{\lambda_1, \ldots, \lambda_r\}$ and $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_r\}$, respectively, the CPEP can be further upper bounded by

$$P(\boldsymbol{x}^c \to \boldsymbol{x}^e|\boldsymbol{h}) \leq \frac{1}{2}\exp\left(-\frac{\gamma_s\sum_{i=1}^r \lambda_i|\bar{h}_i|^2}{4}\right), \qquad (22)$$

where $\bar{h}_i = \langle\boldsymbol{h}, \boldsymbol{\mu}_i\rangle$, $\forall i \in [1, r]$, and we have $\bar{h}_i \sim \mathcal{CN}(v_i, 1/P)$ with $v_i = \langle\mathbb{E}[\boldsymbol{h}], \boldsymbol{\mu}_i\rangle$. It can be readily shown that $|\bar{h}_i|$ are Rician distributed variables having the probability density function (PDF) of [23]

$$p(|\bar{h}_i|) = 2P|\bar{h}_i|\exp(-P|\bar{h}_i|^2 - P\zeta_i)I_0(2P|\bar{h}_i|\sqrt{\zeta_i}), \quad (23)$$

where $\zeta_i = |v_i|^2$ is the Rician factor. By averaging (22) with respect to $|\bar{h}_i|$, the UPEP can be formulated as [23]

$$P(\boldsymbol{x}^c \to \boldsymbol{x}^e) \leq \frac{1}{2}\prod_{i=1}^r \frac{1}{1 + \frac{\lambda_i\gamma_s}{4P}}\exp\left(-\frac{\frac{\zeta_i\lambda_i\gamma_s}{4P}}{1 + \frac{\lambda_i\gamma_s}{4P}}\right). \quad (24)$$

Under the assumption of Rayleigh fading associated with $\zeta_i = 0$, $|h_i|$ follows the Rayleigh distribution. Hence, (24) can be expressed as

$$P(\boldsymbol{x}^c \to \boldsymbol{x}^e) \leq \frac{1}{2\prod_{i=1}^r(1 + \lambda_i\gamma_s/4P)}. \qquad (25)$$

In the sufficiently high SNR regime ($\gamma_s \gg 1$), the UPEP can be formulated as

$$P(\boldsymbol{x}^c \to \boldsymbol{x}^e) \leq \frac{1}{2\prod_{i=1}^r \lambda_i}\left(\frac{\gamma_s}{4P}\right)^{-r}. \qquad (26)$$

Finally, based on the union bounding technique, the BER of OTSM can be approximately written as

$$P_e \approx \frac{1}{L_b2^{L_b}}\sum_{\boldsymbol{x}^c}\sum_{\boldsymbol{x}^e} P(\boldsymbol{x}^c \to \boldsymbol{x}^e)e(\boldsymbol{x}^c, \boldsymbol{x}^e), \qquad (27)$$

where $e(\boldsymbol{x}^c, \boldsymbol{x}^e)$ denotes the corresponding number of different bits between $\boldsymbol{x}^c$ and $\boldsymbol{x}^e$. For high SNR values, the upper bound of the BER can be further simplified based on (26) and (27) as

$$P_e \leq \frac{1}{2L_b2^{L_b}}\sum_{\boldsymbol{x}^c}\sum_{\boldsymbol{x}^e} \frac{1}{\prod_{i=1}^r \lambda_i}\left(\frac{\gamma_s}{4P}\right)^{-r} e(\boldsymbol{x}^c, \boldsymbol{x}^e). \quad (28)$$

B. Imperfect Channel Estimation Conditions

By taking realistic channel estimation imperfections into account, the channel coefficient vector $\boldsymbol{h}$ of (15) can be rewritten as [24]

$$\tilde{\boldsymbol{h}} = \boldsymbol{h} + \boldsymbol{e}_h, \qquad (29)$$

where $\boldsymbol{e}_h$ represents the channel estimation error whose elements obeys $\mathcal{CN}(0, \sigma_h^2)$ with the variance $0 \leq \sigma_h^2 < 1$. Under this condition, the input-output relationship of (15) can be formulated as

$$\boldsymbol{y} = \boldsymbol{\Phi}(\boldsymbol{x})\tilde{\boldsymbol{h}} + \boldsymbol{\Phi}(\boldsymbol{x})(\boldsymbol{h} - \tilde{\boldsymbol{h}}) + \bar{\boldsymbol{n}} = \boldsymbol{\Phi}(\boldsymbol{x})\tilde{\boldsymbol{h}} + \tilde{\boldsymbol{n}}, \quad (30)$$

where we have $\tilde{\boldsymbol{n}} \triangleq \boldsymbol{\Phi}(\boldsymbol{x})(\boldsymbol{h} - \tilde{\boldsymbol{h}}) + \bar{\boldsymbol{n}} = -\boldsymbol{\Phi}(\boldsymbol{x})\boldsymbol{e}_h + \bar{\boldsymbol{n}}$. Therefore, similar to the derivation of (19) and (20), the corresponding CPEP can be expressed as

$$\begin{aligned} P(\boldsymbol{x}^c \to \boldsymbol{x}^e|\tilde{\boldsymbol{h}}) &= P\left[||\boldsymbol{y} - \boldsymbol{\Phi}(\boldsymbol{x}^e)\tilde{\boldsymbol{h}}||_2^2 \leq ||\boldsymbol{y} - \boldsymbol{\Phi}(\boldsymbol{x}^c)\tilde{\boldsymbol{h}}||_2^2\right] \\ &= P[2\Re\left\{\boldsymbol{y}^H\boldsymbol{\Phi}(\boldsymbol{x}^e)\tilde{\boldsymbol{h}}\right\} - 2\Re\left\{\boldsymbol{y}^H\boldsymbol{\Phi}(\boldsymbol{x}^c)\tilde{\boldsymbol{h}}\right\} \\ &\geq ||\boldsymbol{\Phi}(\boldsymbol{x}^e)\tilde{\boldsymbol{h}}||_2^2 - ||\boldsymbol{\Phi}(\boldsymbol{x}^c)\tilde{\boldsymbol{h}}||_2^2] \\ &= P\left[\Re\{\tilde{\boldsymbol{n}}^H\boldsymbol{\Phi}(\boldsymbol{e})\tilde{\boldsymbol{h}}\} \geq \frac{\left\|\boldsymbol{\Phi}(\boldsymbol{e})\tilde{\boldsymbol{h}}\right\|_2^2}{2}\right]. \quad (31) \end{aligned}$$

It can be readily observed that $\Re\{\tilde{\boldsymbol{n}}^H\boldsymbol{\Phi}(\boldsymbol{e})\tilde{\boldsymbol{h}}\}$ has a zero mean and variance of

$$Var\left[\Re\{\tilde{\boldsymbol{n}}^H\boldsymbol{\Phi}(\boldsymbol{e})\tilde{\boldsymbol{h}}\}\right] = \frac{\sigma_h^2\left\|\boldsymbol{\Phi}(\boldsymbol{x}^c)^H\boldsymbol{\Phi}(\boldsymbol{e})\tilde{\boldsymbol{h}}\right\|_2^2 + N_0\left\|\boldsymbol{\Phi}(\boldsymbol{e})\tilde{\boldsymbol{h}}\right\|_2^2}{2}. \quad (32)$$

Therefore, the CPEP of (31) can be rewritten as

$$P(\boldsymbol{x}^c \to \boldsymbol{x}^e|\tilde{\boldsymbol{h}})$$

$$= Q\left(\frac{\left\|\boldsymbol{\Phi}(\boldsymbol{e})\tilde{\boldsymbol{h}}\right\|_2^2}{\sqrt{2\sigma_h^2\left\|\boldsymbol{\Phi}(\boldsymbol{x}^c)^H\boldsymbol{\Phi}(\boldsymbol{e})\tilde{\boldsymbol{h}}\right\|_2^2 + 2N_0\left\|\boldsymbol{\Phi}(\boldsymbol{e})\tilde{\boldsymbol{h}}\right\|_2^2}}\right). \quad (33)$$

Explicitly, the UPEP can be expressed as $P(\boldsymbol{x}^c \to \boldsymbol{x}^e) = \mathbb{E}_{\tilde{\boldsymbol{h}}}\left[P(\boldsymbol{x}^c \to \boldsymbol{x}^e|\tilde{\boldsymbol{h}})\right]$, and $\tilde{\boldsymbol{h}}$ has the multivariate complex Gaussian PDF of

$$f(\tilde{\boldsymbol{h}}) = \frac{\pi^{-P}}{\det(\tilde{\boldsymbol{\Psi}})}\exp\left(-\tilde{\boldsymbol{h}}^H\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{h}}\right), \quad (34)$$

where $\tilde{\boldsymbol{\Psi}} = \mathbb{E}\left\{\tilde{\boldsymbol{h}}\tilde{\boldsymbol{h}}^H\right\} = \boldsymbol{\Psi} + \sigma_h^2\boldsymbol{I}_P$ and $\boldsymbol{\Psi} = \mathbb{E}\left\{\boldsymbol{h}\boldsymbol{h}^H\right\}$. However, given the complex structure of (33), calculating the UPEP directly is intractable. In the case of a constant envelope K-ary modulation constellation $\mathcal{A}$, it can be readily shown that $\left\|\boldsymbol{\Phi}(\boldsymbol{x}^c)^H\boldsymbol{\Phi}(\boldsymbol{e})\tilde{\boldsymbol{h}}\right\|_2^2 \le \left\|\boldsymbol{\Phi}(\boldsymbol{e})\tilde{\boldsymbol{h}}\right\|_2^2$. Therefore, the upper-bound of the CPEP can be formulated as

$$P(\boldsymbol{x}^c \to \boldsymbol{x}^e|\tilde{\boldsymbol{h}}) = Q\left(\sqrt{\frac{\left\|\boldsymbol{\Phi}(\boldsymbol{e})\tilde{\boldsymbol{h}}\right\|_2^2}{2\sigma_h^2 + 2N_0}}\right). \quad (35)$$

Based on (34), (35) and the Chernoff bound $Q(x) \le \frac{1}{2}\exp(-\frac{x^2}{2})$ [23], the UPEP upper-bound of the generalized OTSM system relying on realistic imperfect channel estimation can be derived as

$$P(\boldsymbol{x}^c \to \boldsymbol{x}^e) \le \frac{\pi^{-P}}{2\det(\tilde{\boldsymbol{\Psi}})}\int_{\tilde{\boldsymbol{h}}}\exp\left(-\tilde{\boldsymbol{h}}^H\left[\tilde{\boldsymbol{\Psi}} + \kappa\boldsymbol{\Theta}\right]\tilde{\boldsymbol{h}}\right)d\tilde{\boldsymbol{h}}$$

$$= \frac{1}{2\det(\boldsymbol{I}_P + \kappa\tilde{\boldsymbol{\Psi}}\boldsymbol{\Theta})}, \quad (36)$$

where $\kappa = 1/(4\sigma_h^2 + 4N_0)$ and $\boldsymbol{\Theta} = \boldsymbol{\Phi}(\boldsymbol{e})^H\boldsymbol{\Phi}(\boldsymbol{e})$. Finally, the BER upper bound under channel estimation imperfections can be calculated based on (27).

IV. Detection Algorithms for ZP-OTSM

In this section, we first discuss the optimum maximum *a posteriori* (MAP) detector. Then, the family of linear detectors is introduced. Moreover, the conventional AMP detector is detailed. Furthermore, we propose our low-complexity VAMP-EM detector based on a factor graph. Finally, the complexity of different detectors is discussed.

A. Maximum a posteriori Detection

In general, the optimal MAP detector of the OTSM system described in Section II relies on maximizing the *a posteriori* probability of $\boldsymbol{x}$, given the received symbol vector $\boldsymbol{y}$ as shown in (10), which is formulated as $\hat{\boldsymbol{x}}^{\text{MAP}} = \arg\max_{\boldsymbol{x}\in\mathcal{A}^{MN}}\{p(\boldsymbol{x}|\boldsymbol{y})\}$. Upon assuming that the mapping processes among different symbols are independent and equiprobable, the MAP detector can also be formulated as the ML detector equivalently of

$$\hat{\boldsymbol{x}}^{\text{ML}} = \arg\min_{\boldsymbol{x}\in\mathcal{A}^{MN}}\left\{\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|_2^2\right\}. \quad (37)$$

Nevertheless, the complexity of the optimum ML detector is given by $\mathcal{O}(K^{MN})$, which is excessive for high values of MN. To mitigate this problem, it is crucial to design efficient detectors for large-scale OTSM systems.

B. Linear Minimum Mean Square Error Detector

According to (10), the linear minimum mean square error (LMMSE) detector can be readily employed for recovering $\boldsymbol{x}$, yielding,

$$\hat{\boldsymbol{x}}^{\text{LMMSE}} = (\boldsymbol{H}^H\boldsymbol{H} + \frac{1}{\gamma_s}\boldsymbol{I}_{MN})^{-1}\boldsymbol{H}^H\boldsymbol{y}. \quad (38)$$

As shown in Fig. 3 (b), the block circulant property cannot be satisfied by the ZP-OTSM system. Hence, the complexity of the LMMSE detector is still in the order of $\mathcal{O}(M^3N^3)$, which is consistent with the observations in [25].

C. Conventional AMP Detector

The AMP algorithm is initially proposed based on loopy BP amalgamated with Gaussian and Taylor series approximations, whose message passing process is conceived based on a scalar-valued factor graph [15]. Given the DS-domain input-output relationship in (10), the conventional AMP detector is formally stated in Algorithm 1. The AMP detector includes symbol denoising (SD) and LMMSE estimation (LE).

Given the modulation scheme, the *a prior* information of $\boldsymbol{x}$ can be written as $P(x(j) = a_k) = 1/K$ for $j = 0, \ldots, J-1$ and $k = 1, \ldots, K$. The AMP algorithm quantities $\{\boldsymbol{r}, \boldsymbol{x}\}$ obey

$$r(j) = x(j) + \omega(j), \quad (39)$$

where $\omega(j) \sim \mathcal{CN}[0, v_r(j)]$, and it can be observed that $\boldsymbol{r}$ is viewed as the $\boldsymbol{v_r}$-variance AWGN-contaminated version of the true signal $\boldsymbol{x}$. In the SD part, lines 5-8 of Algorithm 1 are leveraged to compute the *a posteriori* mean $x(j)$ and variance $v_x(j)$ of the elements in $\boldsymbol{x}$. Moreover, the symbol denoising results in $\hat{\boldsymbol{x}}$ can be regarded as an AWGN-contaminated version of the true symbol $\boldsymbol{x}$, which is given by

$$\hat{x}(j) = x(j) + \varpi(j), \quad (40)$$

where $\varpi(j) \sim \mathcal{CN}[0, v_x(j)]$. In the LE part, lines 10-15 of Algorithm 1 are directly given by the AMP algorithm to calculate $r(j)$. The term $\boldsymbol{v}_r^{t+1}\cdot\boldsymbol{H}^H\boldsymbol{s}^{t+1}$ in line 15 of Algorithm 1 is the Onsager term of [15] that asymptotically decouples the denoising part and the LMMSE estimation part through iterations, i.e., eliminates the correlation between these two parts.

It should be noted that the AMP detector has two limitations. First, AMP is designed for accommodating zero-mean i.i.d Gaussian sensing matrices [15]. Bearing in mind the specific structure of the DS-domain channel matrix $\boldsymbol{H}$ shown in Fig. 3 (b), the performance of the AMP detector may degrade significantly or even diverge. Moreover, the AMP detector proceeds iteratively with the known noise variance $1/\gamma_n$, which is however rarely known in practice. To alleviate the above issues, we propose the VAMP-EM detector in the next subsection.

Algorithm 1 AMP Detector

Require: $\boldsymbol{y}$, $\boldsymbol{H}$ and γ_n.
 1: **Preparation**: Set a maximum iterations number T_{AMP} and an error tolerance parameter ϵ.
 2: **Initialize** $\boldsymbol{s}^{(1)} = \boldsymbol{0}$, $\boldsymbol{v}_r^{(1)} = \boldsymbol{1}$ and $\boldsymbol{r}^{(1)} = \boldsymbol{0}$.
 3: **for** $t = 1$ to T_{AMP} **do**
 4:     //Symbol denoising
 5:     $\forall j, k : \xi_{j,k}^t = \exp\left[-|a_k - r^t(j)|^2 / v_r(j)^t\right]$
 6:     $\forall j, k : \beta_{j,k} = \xi_{j,k}^t / \sum_{k=1}^{K} \xi_{j,k}^t$
 7:     $\forall j : \hat{x}^t(j) = \sum_{k=1}^{K} a_k \beta_{j,k}$
 8:     $\forall j : v_x^t(j) = \sum_{k=1}^{K} \beta_{j,k} |a_k - \hat{x}^t(j)|^2$
 9:     //LMMSE estimation
10:     $\boldsymbol{v}_p = |\boldsymbol{H}|^2 \boldsymbol{v}_x^t$
11:     $\boldsymbol{p} = \boldsymbol{H}\hat{\boldsymbol{x}}^t - \boldsymbol{v}_p \cdot \boldsymbol{s}^t$
12:     $\boldsymbol{v}_s = \boldsymbol{1} \cdot / \left(\boldsymbol{v}_p + \gamma_n^{-1}\boldsymbol{1}\right)$
13:     $\boldsymbol{s}^{t+1} = \boldsymbol{v}_s \cdot (\boldsymbol{y} - \boldsymbol{p})$
14:     $\boldsymbol{v}_r^{t+1} = \boldsymbol{1} \cdot / |\boldsymbol{H}^H|^2 \boldsymbol{v}_s$
15:     $\boldsymbol{r}^{t+1} = \hat{\boldsymbol{x}}^t + \boldsymbol{v}_r^{t+1} \cdot \boldsymbol{H}^H \boldsymbol{s}^{t+1}$
16:     **if** $\|\hat{\boldsymbol{x}}^{t+1} - \hat{\boldsymbol{x}}^t\|_2^2 < \epsilon \|\hat{\boldsymbol{x}}^t\|_2^2$
17:         **break**
18: **end for**
19: **return** $\hat{\boldsymbol{x}}$.

D. VAMP-EM Detector

The VAMP algorithm is derived from a non-loopy factor graph having vector-valued nodes, which is different from AMP [17]. Hence, VAMP achieves better MSE performance than AMP in the face of ill-conditioned sensing matrices. By intrinsically amalgamating the VAMP and the EM algorithms, we propose the VAMP-EM detector summarized in Algorithm 2, which includes two inner iterations and one outer iterations. Specifically, we consider three aspects, which are detailed below.
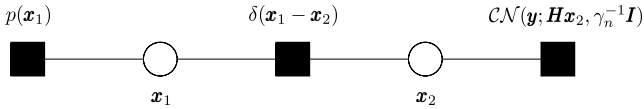

Fig. 4. Illustration of the VAMP-EM factor graph

1) VAMP Algorithm: We first factorise the joint likelihood function of $\boldsymbol{y}$ and $\boldsymbol{x}$ formulated as $p(\boldsymbol{y}, \boldsymbol{x}) = p(\boldsymbol{x})\mathcal{CN}(\boldsymbol{y}; \boldsymbol{Hx}, \boldsymbol{I}/\gamma_n)$. As shown in Fig. 4, based on the derivation of the VAMP illustrated in [17], the transmitted symbol vector $\boldsymbol{x}$ is split into two variables $\boldsymbol{x}_1 = \boldsymbol{x}_2$. Hence, the equivalent factorization of $p(\boldsymbol{y}, \boldsymbol{x})$ can be further expressed as $p(\boldsymbol{y}, \boldsymbol{x}_1, \boldsymbol{x}_2) = p(\boldsymbol{x}_1)\delta(\boldsymbol{x}_1 - \boldsymbol{x}_2)\mathcal{CN}(\boldsymbol{y}; \boldsymbol{Hx}_2, \boldsymbol{I}/\gamma_n)$. In the tth outer loop iteration, the message passing process commences with $\mu_{\delta \to \boldsymbol{x}_1}(\boldsymbol{x}_1) = \mathcal{CN}\left[\boldsymbol{x}_1; \boldsymbol{r}_1^t, (\gamma_1^t \boldsymbol{I})^{-1}\right]$. Then the approximate belief of $\boldsymbol{x}_1$ can be calculated as $b_{\text{app}}(\boldsymbol{x}_1) = \mathcal{CN}\left[\boldsymbol{x}_1; \hat{\boldsymbol{x}}_1^t, (\eta_1^t)^{-1}\boldsymbol{I}\right]$.

In the SD part, based on the statistics of the modulation scheme, the probability density function of $\boldsymbol{x}$ can be formulated as $p(\boldsymbol{x}) = \prod_{j=1}^{MN} p(x_j) = \prod_{j=1}^{MN} \frac{1}{K} \sum_{k=1}^{K} \delta(x_j - a_k)$. Similar to AMP, we obtain the scalar-valued *a posteriori* mean $\hat{x}_1(j)$ and variance $v_{x_1}(j)$, yielding lines 6-9 of Algorithm 2. Thus the average variance can be written as $1/\eta_1^\tau = \frac{1}{MN} \sum_{j=1}^{MN} v_{x_1}(j)$, yielding line 10 of Algorithm 2. Hence we obtain $b_{\text{app}}(\boldsymbol{x}_1)$. In the classic VAMP algorithm, under the assumption that $\boldsymbol{H}$ is right-rotationally invariant, both $\hat{\boldsymbol{x}}_1$ and

$\hat{\boldsymbol{x}}_2$ can be obtained by LMMSE estimation. Then the outputs of the SD and LE parts obey [17]

$$\boldsymbol{r}_1^t = \boldsymbol{x} + \mathcal{CN}(0, \boldsymbol{I}/\gamma_1^t),$$
$$\boldsymbol{x} = \boldsymbol{r}_2^t + \mathcal{CN}(0, \boldsymbol{I}/\gamma_2^t). \tag{41}$$

Hence, $\boldsymbol{r}_1^t$ may be viewed as a $1/\gamma_1^t$-variance AWGN-contaminated version of the true signal $\boldsymbol{x}$, and the true symbol vector $\boldsymbol{x}$ can be regarded as a $1/\gamma_2^t$-variance AWGN-contaminated version of $\boldsymbol{r}_2^t$. Based on the classic message passing rules of VAMP [17], the message $\mu_{\boldsymbol{x}_1 \to \delta}(\boldsymbol{x}_1) = \mathcal{CN}\left[\boldsymbol{x}_1; \boldsymbol{r}_2^t, (\gamma_2^t)^{-1}\boldsymbol{I}\right]$ can be formulated as

$$\mu_{\boldsymbol{x}_1 \to \delta}(\boldsymbol{x}_1) = \mathcal{CN}\left[\boldsymbol{x}_1; \hat{\boldsymbol{x}}_1^t, (\eta_1^t)^{-1}\boldsymbol{I}\right] / \mathcal{CN}(\boldsymbol{x}_1; \boldsymbol{r}_1^t, (\gamma_1^t)^{-1}\boldsymbol{I})$$
$$\propto \mathcal{CN}\left[\boldsymbol{x}_1; (\hat{\boldsymbol{x}}_1^t \eta_1^t - \boldsymbol{r}_1^t \gamma_1^t)/(\eta_1^t - \gamma_1^t), (\eta_1^t - \gamma_1^t)^{-1}\right]. \tag{42}$$

Therefore, we can obtain $\boldsymbol{r}_2^t = (\eta_1^t \hat{\boldsymbol{x}}_1^t - \gamma_1^t \boldsymbol{r}_1^t)/\gamma_2^t$ and $\gamma_2^t = \eta_1^t - \gamma_1^t$, yielding lines 14-15 of Algorithm 2. It can be readily shown in Fig. 4 that the message $\mu_{\boldsymbol{x}_1 \to \delta}(\boldsymbol{x}_1) = \mathcal{CN}\left[\boldsymbol{x}_1; \boldsymbol{r}_2^t, (\gamma_2^t)^{-1}\boldsymbol{I}\right]$ remains constant after passing through the node δ, hence we have $\mu_{\delta \to \boldsymbol{x}_2}(\boldsymbol{x}_2) = \mu_{\boldsymbol{x}_1 \to \delta}(\boldsymbol{x}_1) = \mathcal{CN}\left[\boldsymbol{x}_1; \boldsymbol{r}_2^t, (\gamma_2^t)^{-1}\boldsymbol{I}\right]$.

In the LE part, according to the factor graph message passing rules [17], the approximate belief of the node $\boldsymbol{x}_2$ in Fig. 4 can be expressed as $b_{\text{app}}(\boldsymbol{x}_2) = \mathcal{CN}\left[\boldsymbol{x}_2, \hat{\boldsymbol{x}}_2^t, (\gamma_2^t)^{-1}\boldsymbol{I}\right]$, where $\hat{\boldsymbol{x}}_2^t = \mathbb{E}[\boldsymbol{x}_2 | b_{\text{sp}}(\boldsymbol{x}_2)]$ and $\eta_2^t = \langle \text{diag}(\text{Cov}[\boldsymbol{x}_2 | b_{\text{sp}}(\boldsymbol{x}_2)])\rangle^{-1}$, the sum-product (SP) belief concerning $\boldsymbol{x}_2$ can be formulated as $b_{\text{sp}}(\boldsymbol{x}_2) \propto \mathcal{CN}\left[\boldsymbol{x}_2; \boldsymbol{r}_2^t, (\eta_2^t)^{-1}\boldsymbol{I}\right]\mathcal{CN}\left[\boldsymbol{y}; \boldsymbol{Hx}_2, (\gamma_n)^{-1}\boldsymbol{I}\right]$. Consequently, the conditional-mean and covariance of $b_{\text{app}}(\boldsymbol{x}_2)$ can be formulated as [17]

$$\hat{\boldsymbol{x}}_2^t = \left(\gamma_n^\tau \boldsymbol{H}^H \boldsymbol{H} + \gamma_2^\tau \boldsymbol{I}\right)^{-1}\left(\gamma_n^\tau \boldsymbol{H}^H \boldsymbol{y} + \gamma_2^\tau \boldsymbol{r}_2^\tau\right), \tag{43}$$

$$(\eta_2^t)^{-1}\boldsymbol{I} = \left(\gamma_n^\tau \boldsymbol{H}^H \boldsymbol{H} + \gamma_2^\tau \boldsymbol{I}\right)^{-1}. \tag{44}$$

Hence, upon exploiting

$$\boldsymbol{g}_2(\boldsymbol{r}_2^\tau, \gamma_2^\tau) = \left(\gamma_n^\tau \boldsymbol{H}^H \boldsymbol{H} + \gamma_2^\tau \boldsymbol{I}\right)^{-1}\left(\gamma_n^\tau \boldsymbol{H}^H \boldsymbol{y} + \gamma_2^\tau \boldsymbol{r}_2^\tau\right), \tag{45}$$

$$\langle \boldsymbol{g}_2'(\boldsymbol{r}_2^\tau, \gamma_2^\tau)\rangle = \gamma_2^\tau \text{tr}\left\{\left(\gamma_n^\tau \boldsymbol{H}^H \boldsymbol{H} + \gamma_2^\tau \boldsymbol{I}\right)^{-1}\right\} / MN, \tag{46}$$

lines 18-20 of Algorithm 2 have been interpreted.

In order to avoid the matrix inversion process in (45) and (46), we apply the singular value decomposition (SVD) to the DS-domain DT channel matrix, yielding $\boldsymbol{H} = \boldsymbol{USV}^H$, where we have $\boldsymbol{S} = \text{diag}[s_1, \ldots, s_{MN}]$ and $R = \text{rank}(\boldsymbol{H})$. Hence, $\boldsymbol{g}_2(\boldsymbol{r}_2^\tau, \gamma_2^\tau)$ and $\langle \boldsymbol{g}_2'(\boldsymbol{r}_2^\tau, \gamma_2^\tau)\rangle$ may be written as

$$\boldsymbol{g}_2(\boldsymbol{r}_2^\tau, \gamma_2^\tau) = \boldsymbol{V}\Xi^\tau(\bar{\boldsymbol{y}} + \gamma_2^\tau \boldsymbol{V}^H \boldsymbol{r}_2^\tau), \tag{47}$$

$$\langle \boldsymbol{g}_2'(\boldsymbol{r}_2^\tau, \gamma_2^\tau)\rangle = \frac{1}{MN} \sum_{j=1}^{MN} \frac{\gamma_2^\tau}{\gamma_n |s_j|^2 + \gamma_2^\tau}, \tag{48}$$

where $\bar{\boldsymbol{y}} = \gamma_n \boldsymbol{S}^H \boldsymbol{U}^H \boldsymbol{y}$ and $\Xi^\tau \in \mathbb{R}^{MN \times MN}$ denotes a diagonal matrix associated with $\Xi^\tau(j, j) = (\gamma_n |s_j|^2 + \gamma_2^\tau)^{-1}$. Then we obtain the *a posteriori* mean symbol vector $\hat{\boldsymbol{x}}_2^t$ and the variance $1/\eta_2^t$. It maybe readily shown that the message

$\mu_{\boldsymbol{x}_2 \to \delta}(\boldsymbol{x}_2)$ can be formulated as [17]

$$
\begin{aligned}
\mu_{\boldsymbol{x}_2 \to \delta}(\boldsymbol{x}_2) &= \mathcal{CN}\left[\boldsymbol{x}_2; \boldsymbol{r}_1^{t+1}, (\gamma_1^{t+1})^{-1}\boldsymbol{I}\right] \\
&= \mathcal{CN}\left[\boldsymbol{x}_2; \hat{\boldsymbol{x}}_2^t, (\eta_2^t)^{-1}\boldsymbol{I}\right] / \mathcal{CN}\left[\boldsymbol{x}_2; \boldsymbol{r}_2^t, (\gamma_2^t)^{-1}\right].
\end{aligned}
\tag{49}
$$

Consequently, we can obtain $\boldsymbol{r}_1^{t+1} = (\eta_2^t \hat{\boldsymbol{x}}_2^t - \gamma_2^t \boldsymbol{r}_2^t)/(\eta_2^t - \gamma_2^t)$ and $\gamma_1^{t+1} = \eta_2^t - \gamma_2^t$. It should be noted that the DS-domain channel matrix $\boldsymbol{H}$ cannot satisfy the right-rotationally invariant property. Hence, for the ill-conditioned $\boldsymbol{H}$, the performance of the VAMP-EM detector may be poor. To circumvent this impediment, a damping strategy is applied to the VAMP-EM detection algorithm [16], [19]. This justifies lines 25-26 of Algorithm 2, where the damping factor obeys $\theta \in (0, 1]$. In general, the optimal θ can be obtained for a given SNR by searching within the interval $\theta \in (0 : 0.1 : 1]$ based on the minimum BER criteria. Since the message through the node δ remains unchanged, we have $\mu_{\delta \to \boldsymbol{x}_1}(\boldsymbol{x}_1) = \mathcal{CN}\left[\boldsymbol{x}_1; \boldsymbol{r}_1^{t+1}, (\gamma_1^{t+1})^{-1}\right]$. Then the above message passing iterations are repeated with $t \leftarrow t + 1$.

Compared to AMP, we noticed that both the AMP and VAMP include the SD and LE parts, hence the denoising steps in lines 5-8 of Algorithm 1 are equivalent to lines 6-9 of Algorithm 2. But the outputs of the SD part within the AMP detector are the *a posteriori* mean $\boldsymbol{x}$ and variance $\boldsymbol{v_x}$ without Onsager correction, which leads to correlations between $\boldsymbol{r}$ as well as $\boldsymbol{x}$, and to a loss of performance.

Algorithm 2 VAMP-EM Detector

Require: $\boldsymbol{y}$, $\boldsymbol{H} = \boldsymbol{USV}^H$ with $\boldsymbol{S} = \mathrm{diag}[s_1, \ldots, s_{MN}]$ and $R = \mathrm{rank}(\boldsymbol{H}) \leq MN$.
1: **Preparation**: Set a maximum iterations number T, a damping factor θ and an error tolerance parameter ϵ.
2: **Initialize** $\gamma_1^{(1)} = 0$, $1/\gamma_n^{(1)} = ||\boldsymbol{y}||_2^2/MN$ and $\boldsymbol{r}_1^{(1)} = \boldsymbol{0}$.
3: **for** $t = 1$ to T **do**
4: $\quad$ //Symbol denoising
5: $\quad$ **for** $\tau = 1$ to T_1 **do**
6: $\quad\quad$ $\forall j, k : \xi_{j,k}^\tau = \exp\left[-\gamma_1^\tau |a_k - r_1^\tau(j)|^2\right]$
7: $\quad\quad$ $\forall j, k : \beta_{j,k}^\tau = \xi_{j,k}^\tau / \sum_{k=1}^K \xi_{j,k}^\tau$
8: $\quad\quad$ $\forall j : \hat{x}_1^\tau(j) = \sum_{k=1}^K a_k \beta_{j,k}^\tau$
9: $\quad\quad$ $\forall j : v_{x_1}(j) = \sum_{k=1}^K \beta_{j,k}^\tau |a_k - \hat{x}_1^\tau(j)|^2$
10: $\quad\quad$ $1/\eta_1^\tau = \frac{1}{MN} \sum_{j=1}^{MN} v_{x_1}(j)$
11: $\quad\quad$ $1/\gamma_1^{\tau+1} = \frac{1}{MN} ||\hat{\boldsymbol{x}}_1 - \boldsymbol{r}_1^\tau||_2^2 + 1/\eta_1^\tau$
12: $\quad$ **end for**
13: $\quad$ $\hat{\boldsymbol{x}}_1^t = \hat{\boldsymbol{x}}_1^{T_1}$, $1/\eta_1^t = 1/\eta_1^{T_1}$, $1/\gamma_1^t = 1/\gamma_1^{T_1+1}$
14: $\quad$ $\gamma_2^t = \eta_1^t - \gamma_1^t$
15: $\quad$ $\boldsymbol{r}_2^t = (\eta_1^t \hat{\boldsymbol{x}}_1^t - \gamma_1^t \boldsymbol{r}_1^t)/\gamma_2^t$
16: $\quad$ //LMMSE estimation
17: $\quad$ **for** $\tau = 1$ to T_2 **do**
18: $\quad\quad$ $\hat{\boldsymbol{x}}_2^\tau = \boldsymbol{g}_2(\boldsymbol{r}_2^\tau, \gamma_2^\tau)$
19: $\quad\quad$ $\alpha_2^\tau = \langle \boldsymbol{g}_2'(\boldsymbol{r}_2^\tau, \gamma_2^\tau) \rangle$
20: $\quad\quad$ $1/\eta_2^\tau = \alpha_2^\tau / \gamma_2^\tau$
21: $\quad\quad$ $1/\gamma_2^{\tau+1} = \frac{1}{MN} ||\hat{\boldsymbol{x}}_2^\tau - \boldsymbol{r}_2^\tau||_2^2 + 1/\eta_2^\tau$
22: $\quad\quad$ $1/\gamma_n^{\tau+1} = \frac{1}{MN}\left[||\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{x}}_2^\tau||_2^2 + \sum_{k=1}^R \frac{|s_k|^2}{\gamma_n^\tau |s_k|^2 + \gamma_2^{\tau+1}}\right]$
23: $\quad$ **end for**
24: $\quad$ $\hat{\boldsymbol{x}}_2^t = \hat{\boldsymbol{x}}_2^{T_2}$, $1/\eta_2^t = 1/\eta_2^{T_2+1}$, $1/\gamma_2^t = 1/\gamma_2^{T_2+1}$
25: $\quad$ $\gamma_1^{t+1} = (1 - \theta)\gamma_1^t + \theta(\eta_2^t - \gamma_2^t)$
26: $\quad$ $\boldsymbol{r}_1^{t+1} = (1 - \theta)\boldsymbol{r}_1^t + \theta(\eta_2^t \hat{\boldsymbol{x}}_2^t - \gamma_2^t \boldsymbol{r}_2^t)/(\eta_2^t - \gamma_2^t)$
27: $\quad$ **if** $||\hat{\boldsymbol{x}}_1^{t+1} - \hat{\boldsymbol{x}}_1^t||_2^2 < \epsilon ||\hat{\boldsymbol{x}}_1^t||_2^2$
28: $\quad\quad$ **break**
29: **end for**
30: **return** $\hat{\boldsymbol{x}}_1$.

---

2) Learning γ_n: Now we consider the problem of estimating the noise variance $1/\gamma_n$. Given the DS-domain input-output relationship shown in (10), the likelihood function of $\boldsymbol{y}$ can be written as

$$
p(\boldsymbol{y}|\boldsymbol{x}; \gamma_n) = \left(\frac{\gamma_n}{\pi}\right)^{MN} \exp\left[-\gamma_n ||\boldsymbol{y} - \boldsymbol{Hx}||_2^2\right].
\tag{50}
$$

Consequently, the joint probability density function of $\boldsymbol{x}$ and $\boldsymbol{y}$ given γ_n can be formulated as $p(\boldsymbol{x}, \boldsymbol{y}; \gamma_n) = p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x}; \gamma_n)$. Then the EM algorithm is employed to obtain the ML estimate of the unknown hyper-parameters by iteratively minimizing the negative likelihood upper bound and tightening it [26]. The process of the EM algorithm can be formulated as

$$
V\left(\gamma_n; \hat{\gamma}_n^t\right) \triangleq -\mathbb{E}\left[\ln p(\boldsymbol{x}, \boldsymbol{y}; \gamma_n) \mid \boldsymbol{y}; \hat{\gamma}_n^t\right],
\tag{51}
$$

$$
\hat{\gamma}_n^{t+1} = \arg \min_{\gamma_n} V\left(\gamma_n; \hat{\gamma}_n^t\right).
\tag{52}
$$

The generalized EM framework based on Gibbs sampling is detailed in **Appendix A**. To compute (51) and (52), we consider the Gibbs energy function of [27]

$$
J(b, q; \gamma_n) = D_{\mathrm{KL}}(b, \gamma_n) + H(q),
\tag{53}
$$

which is equivalent to (68), when $b = q$. Hence the minimization of the negative likelihood upper bound can be rewritten as

$$
\begin{aligned}
&\hat{\gamma}_n^{t+1} = \arg \min_{\gamma_n} \min_b \max_q J(b, q, \gamma_n); \\
&\text{s.t. } \mathbb{E}[\boldsymbol{x}|b] = \mathbb{E}[\boldsymbol{x}|q], \quad \mathrm{tr}\{\mathrm{Cov}[\boldsymbol{x}|b]\} = \mathrm{tr}\{\mathrm{Cov}[\boldsymbol{x}|q]\}.
\end{aligned}
\tag{54}
$$

According to **Appendix A**, γ_n can be estimated by solving the following optimization:

$$
\hat{\gamma}_n^{t+1} = \arg \max_{\gamma_n} \mathbb{E}\left[\ln p(\boldsymbol{y}|\boldsymbol{x}, \gamma_n) | \boldsymbol{r}_2^t, \gamma_2^t, \hat{\gamma}_n^t\right].
\tag{55}
$$

Specifically, based on the Gaussian likelihood function (50), we obtain $\hat{\gamma}_n^{t+1}$ as

$$
\begin{aligned}
1/\hat{\gamma}_n^{t+1} &= \frac{1}{MN} \mathbb{E}\left[||\boldsymbol{y} - \boldsymbol{Hx}||_2^2 | \boldsymbol{r}_2^t, \gamma_2^t, \gamma_n^t\right] \\
&= \frac{1}{MN}\left[||\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}_2^t||_2^2 + \sum_{k=1}^R \frac{|s_k|^2}{\gamma_n^t |s_k|^2 + \gamma_2^t}\right],
\end{aligned}
\tag{56}
$$

thus interpreting line 22 of Algorithm 2.

3) Adaptive Auto-tuning: Under the condition that the input-output relationship (10) relying on imperfectly estimated noise variance $1/\gamma_n$, the statistical model of VAMP quantities shown in (41) cannot be satisfied. Specifically, in this case, $\boldsymbol{r}_1^t$ is obtained as an AWGN-contaminated version of the true signal $\boldsymbol{x}$ associated with another AWGN precision value based on the state evolution function [17]. Therefore, the values of $\boldsymbol{r}_i^t$ are characterized by γ_i^t imprecisely hence resulting in the imperfect estimation of γ_n. In order to mitigate this problem, inspired by the adaptive VAMP algorithm [18], the re-estimation process of the noise precision $\{\gamma_i^t\}$ is applied to the VAMP-EM detector. Based on (41), the ML estimation of γ_1^t can be formulated as $\gamma_1^t = \arg \max_{\gamma_1} p(\boldsymbol{r}_1^t; \gamma_1)$. By employing the EM algorithm, the above problem can be solved

indexed by τ, yielding

$$\gamma_1^{\tau+1} = \arg \max_{\gamma_1} \mathbb{E}\left[\ln p(\boldsymbol{x}, \boldsymbol{r}_1^\tau; \gamma_1)|\boldsymbol{r}_1^\tau; \gamma_1^\tau\right]$$
$$= \arg \max_{\gamma_1} \mathbb{E}\left[\ln p(\boldsymbol{r}_1^\tau|\boldsymbol{x}; \gamma_1)|\boldsymbol{r}_1^\tau; \gamma_1^\tau\right]. \quad (57)$$

Therefore, the iterations of γ_1^τ can be further expressed as

$$\gamma_1^{\tau+1} = \arg \max_{\gamma_1} \left\{ MN \ln \gamma_1 - \gamma_1 \mathbb{E}\left[||\boldsymbol{x} - \boldsymbol{r}_1^t||_2^2|\boldsymbol{r}_1^t; \gamma_1^\tau\right] \right\}$$
$$= MN \left\{ \mathbb{E}\left[||\boldsymbol{x} - \boldsymbol{r}_1^t||_2^2|\boldsymbol{r}_1^t; \gamma_1^\tau\right] \right\}^{-1}$$
$$= \left\{ \frac{1}{MN} \sum_{j=1}^{MN} \mathbb{E}[|x(j) - r_1(j)|^2|\boldsymbol{r}_1^\tau; \gamma_1^\tau] \right\}^{-1}$$
$$= \left\{ \frac{1}{MN}||\boldsymbol{x} - \boldsymbol{r}_1^\tau||_2^2 + \frac{1}{\eta_1^\tau} \right\}^{-1}. \quad (58)$$

Finally, the re-estimation process of γ_1^t is detailed above, and the re-estimation result of γ_2^t can be attained similarly. Hence, by employing the EM algorithm, we carry out the first T_1-iterations within the inner loop, while the second inner loop contains T_2 iterations, yielding lines 5-12 and 17-23 of Algorithm 2.

E. Complexity Analysis

The complexity of the single-tap detector is on the order of $\mathcal{O}(MN\log_2 M)$ [14]. Moreover, the complexity order of the GS detector is given by $\mathcal{O}(M^2NL + M^2NT_{\mathrm{GS}})$ [14]. Furthermore, the complexity of the unitary AMP (UAMP) detector can be formulated as $\mathcal{O}(M^2NT_{\mathrm{UAMP}} + MNQT_{\mathrm{UAMP}})$ [11].

Based on the analysis in Section IV-C, the SD part of the conventional AMP detector has a complexity order given by $\mathcal{O}(MNQT_{\mathrm{AMP}})$, while the LE part's complexity is on the order of $\mathcal{O}(MN^2LT_{\mathrm{AMP}})$, which takes advantage of the sparse structure of $\boldsymbol{H}$. Hence, the overall complexity of the AMP detector can be expressed as $\mathcal{O}(MNQT_{\mathrm{AMP}} + MN^2LT_{\mathrm{AMP}})$.

Finally, we consider the VAMP-EM detector of Section IV-D. Similarly, the VAMP-EM detector complexity of the SD part in every iteration is given by $\mathcal{O}(MNQT_1)$, while that of the LE part in every iteration is on the order of $\mathcal{O}(M^2N^2T_2)$. Therefore, the overall complexity of the VAMP-EM detector can be expressed as $\mathcal{O}[(MNQT_1 + M^2N^2T_2)T]$.

V. TURBO RECEIVER OF THE CODED OTSM SYSTEM

In this section, we first discuss the LDPC-coded OTSM system, then detail the structure of both the AMP and VAMP-EM turbo receivers.

We consider the LDPC-coded OTSM system shown in Fig. 1. The turbo receiver includes a soft-decision detector, deinterleaver, and soft LDPC decoder. In the receiver, the soft-information of detected symbols is iteratively exchanged between these three blocks in the form of extrinsic log-likelihood ratios (LLRs). The extrinsic LLRs output by the

symbol-to-bit converter (SBC) can be expressed as [28]

$$L_e^E(c_j(n)) = \ln \frac{P(c_j(n)=0|\boldsymbol{y})}{P(c_j(n)=1|\boldsymbol{y})} - L_a^D(c_j(n))$$
$$= \ln \frac{\sum\limits_{\forall a_k: s_k(n)=0} \varrho_{j,k} \prod\limits_{\forall n':n'\neq n} P(c_j(n')=s_k(n'))}{\sum\limits_{\forall a_k: s_k(n)=1} \varrho_{j,k} \prod\limits_{\forall n':n'\neq n} P(c_j(n')=s_k(n'))}, \quad (59)$$

where $\varrho_{j,k} = \exp\left(-\frac{|\chi(j)-a_k|^2}{\sigma(j)}\right)$ and $L_a^D(c_j(n)) = P(c_j(n)=0)/P(c_j(n)=1)$ represents the *a priori* LLRs obtained from the interleaver, for $j = 0, \ldots, J-1$ and $n = 1, \ldots, \log_2 K$. Furthermore, $\chi(j)$ and $\sigma(j)$ represent the extrinsic mean and variance of $x(j)$, respectively. In the AMP turbo receiver, the output of Algorithm 1 can be modified as [11]

$$\chi(j) = r(j), \quad \sigma(j) = v_r(j). \quad (60)$$

Let us now consider Algorithm 2. As shown in Fig. 4, based on the derivation of the VAMP algorithm [17] and lines 25-26 of Algorithm 2, it can be found that $\boldsymbol{r}_1$ and γ_1 are attained based on the message passed from the variable node $\boldsymbol{x}_2$ to the factor node $\delta(\boldsymbol{x}_1 - \boldsymbol{x}_2)$, which is independent of the *a priori* message gleaned from the variable node $\boldsymbol{x}_1$. Hence, $\boldsymbol{r}_1$ and γ_1 can be regarded as the the output of Algorithm 2, yielding

$$\chi(j) = r_1(j), \quad \sigma(j) = \gamma_1. \quad (61)$$

Moreover, for the symbol denosing parts of turbo receivers, the *a priori* symbol input probability matrix $\boldsymbol{P}$ can be provided based on $L_a^D(c_j(n))$ from the bit-to-symbol converter (BSC). Therefore, line 5 of Algorithm 1 is modified as

$$\xi_{j,k}^t = P(x(j) = a_k) \exp\left[-|a_k - r^t(j)|^2/v_r(j)^t\right], \quad (62)$$

and line 6 of Algorithm 2 can be amended as

$$\xi_{j,k}^\tau = P(x(j) = a_k) \exp\left[-\gamma_1^\tau|a_k - r_1^\tau(j)|^2\right]. \quad (63)$$

where $P(x(j) = a_k)$ represent the elements of $\boldsymbol{P}$. Furthermore, the probability $P(x(j) = a_k)$ can be formulated as [28]

$$P(x(j) = a_k) = \prod_{n=1}^{\log_2 K} P(c_j(n) = s_k(n))$$
$$= \prod_{n=1}^{\log_2 K} 1/2 \left[1 + \tilde{s}_k(n) \tanh(L_a^D(c_j(n)/2))\right], \quad (64)$$

where

$$\tilde{s}_k(n) = \begin{cases} +1, & s_k(n) = 0 \\ -1, & s_k(n) = 1. \end{cases} \quad (65)$$

VI. NUMERICAL RESULTS

In this section, simulation results are provided for characterizing the overall attainable performance of the proposed detectors in both uncoded and coded OTSM systems.

We first consider generalized OTSM systems having the parameters of Table II. Specifically, the channel coefficients

TABLE II
SIMULATION PARAMETERS FOR BER PERFORMANCE ANALYSIS

Parameters	Values
Maximum delay-domain grid index, M	2
No. of OTSM symbol, N	2, 4
Carrier frequency, f_c	4 GHz
Subcarrier spacing, Δf	3.75 kHz
No. of paths, P	4
Maximum normalized delay shift index, $l_{\max}$	1
Maximum normalized Doppler shift index, $k_{\max}$	1, 3
Velocity, v	506.25, 800 km/h

obey $h_i \sim \mathcal{CN}(0, 1/P)$. Corresponding to $N = 2$ and $N = 4$, the maximum speeds are set as $v = 506.25$ km/h and $v = 800$ km/h, respectively. The maximum normalized delay and Doppler indices are $l_{\max} = M-1$ and $k_{\max} = N-1$ [4], where the integer-valued normalized delay and Doppler indices of the ith path are considered as $a_i \in \mathcal{U}[1, l_{\max}]$ ($a_1 = 0$) and $b_i \in \mathcal{U}[-k_{\max}, k_{\max}]$, respectively.
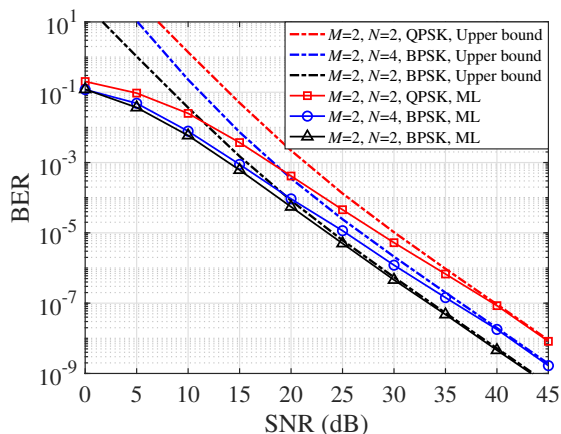
Fig. 5. BER performance of the ML detector and the upper bound of generalized OTSM systems with "$M = 2, N = 2, \text{QPSK}$", "$M = 2, N = 4, \text{BPSK}$" and "$M = 2, N = 2, \text{BPSK}$" based on (28).
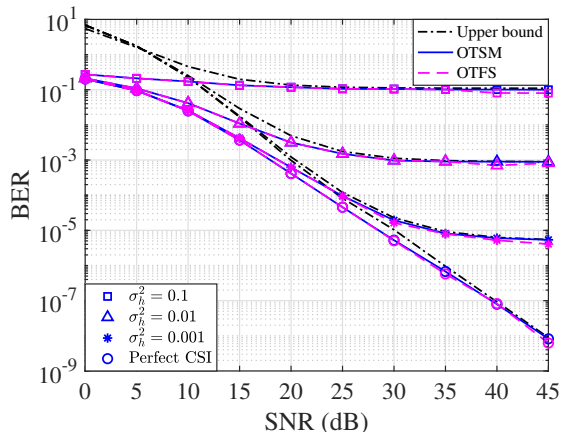
Fig. 6. BER performance of both "$M = 2, N = 2, \text{QPSK}$" OTSM and OTFS systems using ML detector and the upper bound with different values of σ_h^2 based on (27) and (36).

In Fig. 5, the BER performance of the ML detector and the theoretical upper bounds are investigated for parameter settings of "$M = 2, N = 2, \text{QPSK}$", "$M = 2, N = 4, \text{BPSK}$" and "$M = 2, N = 2, \text{BPSK}$". Based on the simulation results of Fig. 5, we have the following observations. Firstly, it is clear that the upper bound becomes very tight as the SNR values increase. Furthermore, for a given modulation scheme,

the OTSM system associated with $MN = 4$ an approximately 3 dB lower SNR at BER=10^{-7} than the $MN = 8$ OTSM system. This is because the value of $e(\boldsymbol{x}^c, \boldsymbol{x}^e)$ becomes higher upon increasing the MN values, leading a higher value of P_e in (27).

Fig. 6 further plots the BER performance of both "$M = 2, N = 2, \text{QPSK}$" OTFS and OTSM systems using the ML detector and the corresponding upper bounds under imperfect channel estimation conditions. The channel estimation variance is set as $\sigma_h^2 = 0.1, 0.01$ and 0.001, respectively. It can be observed from Fig. 6 that, as expected, a lower value of σ_h^2 results in an improved BER performance. Moreover, similar to Fig. 5, the theoretical upper bounds become tight, as the SNRs escalate. Furthermore, OTSM attains a similar BER performance to that of OTFS under all σ_h^2 conditions, which is consistent with the simulations in [14]. Finally, we observe that there are BER error floors at 10^{-1}, 10^{-3} and 5×10^{-6} for $\sigma_h^2 = 0.1, 0.01$ and 0.001, respectively. This is because the channel estimation error dominates the BER performance, even if the value of SNR escalates.

Next, we characterize the BER performance and complexity of the proposed VAMP-EM detectors in both uncoded and LDPC-coded systems, the parameters employed are shown in Table III. Explicitly, we consider generating the channel excess tap delay and impulse response power profiles according to the nine-path extended vehicular A (EVA) channel model [14]. The ith path normalized Doppler shift index is generated upon the Jake's spectrum, i.e. $k_i = k_{\max} \cos(\phi_i)$, where $\phi_i \in \mathcal{U}[0, 2\pi]$.

TABLE III
SIMULATION PARAMETERS FOR CHARACTERIZING THE OVERALL PERFORMANCES OF DETECTORS

Parameters	Values
Maximum delay-domain grid index, M	16
No. of OTSM symbol, N	16
Carrier frequency, f_c	16 GHz
Subcarrier spacing, Δf	60 kHz
Channel model	EVA channel
No. of path, P	9
Length of ZP, L_{ZP}	4
Velocity, v	240, 480 km/h
Damping factor, θ	0.8
Tolerance parameter, ϵ	10^{-10}
Maximum no. of AMP iterations, T_{AMP}	6
Maximum no. of UAMP iterations, T_{UAMP}	10
Maximum no. of VAMP-EM iterations, T	4
Maximum no. of VAMP-EM inner iterations, T_1, T_2	2, 1
Maximum no. of GS iterations, T_{GS}	50
LDPC-coded rate, R	1/2, 3/4

In Fig. 7, the BER performance of the single tap detection of [14], of the AMP, LMMSE, GS [13], UAMP [11] and of the proposed VAMP-EM detectors is investigated for different modulation schemes and relative speeds with integer-valued delay and Doppler shifts. Moreover, the conventional OFDM system associated with LMMSE detection is exploited as a benchmark. We have the following observations based on Fig. 7. Firstly, we observe that the proposed VAMP-EM detector attains a significantly better BER performance than the other detectors. This is because the *a priori* information of the
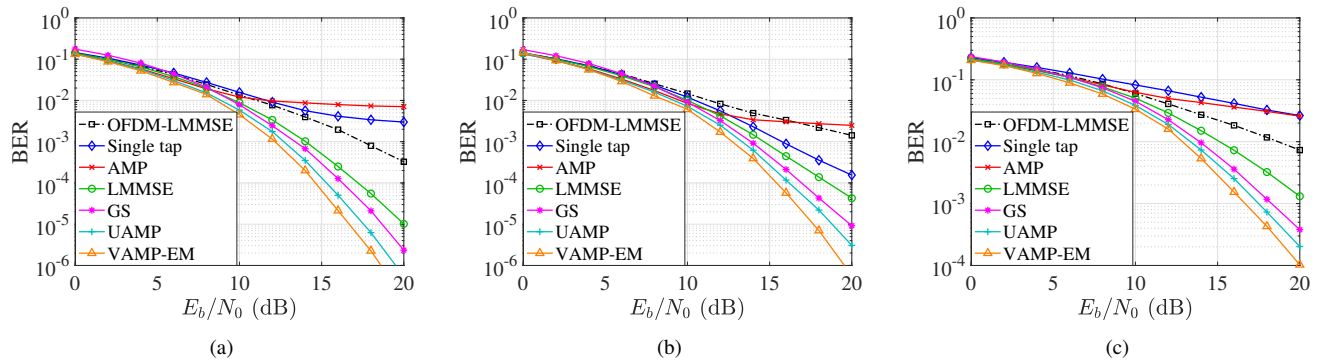
Fig. 7. BER performance of the single tap [14], AMP, LMMSE, GS [13], UAMP [11] and the proposed VAMP-EM detection for the OTSM system with integer-valued delay and Doppler shifts communicating over time-varying EVA channels operating at (a) $(K = 4, v = 480km/h)$ (b) $(K = 4, v = 240km/h)$ (c) $(K = 16, v = 480km/h)$.

modulation scheme is exploited in the VAMP-EM, leading to a BER performance gain. In contrast to the AMP detector, the VAMP-EM detector is derived based on the vector-valued factor graph, making VAMP-EM capable of handling ill-conditioned OTSM channel matrices. Secondly, in the SNR region above 12 dB, the AMP detector has a BER floor. This is because the AMP cannot guarantee convergence in the face of practical sensing matrices. Thirdly, observe in Fig. 7 (a) and Fig. 7 (b) for the BER values of 10^{-5}, and in Fig. 7 (c) for 10^{-3}, that the VAMP-EM detector requires 2.5 dB lower SNR than the GS detector. At BERs of 10^{-5} in Fig. 7 (a) and 10^{-3} in Fig. 7 (c), our VAMP-EM is capable of attaining 1 dB SNR gain compared to the UAMP detector, while this performance gap escalates to about 1.5 dB at a BER of 10^{-5} in Fig. 7 (b). These observations can be explained by the analytical results shown in Section IV-D. Specifically, the VAMP-EM relies on the statistics gleaned from the last iteration and performs LMMSE estimation at every iteration. Moreover, as shown in Fig. 7, when the relative speed is fixed, the smaller the constellation size, the better the BER performance becomes. This is because a high constellation size exhibits a low minimum Euclidean distance between the symbols. Additionally, it should be noted that all the other counterparts are required to know the value of noise variance. Therefore, the simulation results shown in Fig. 7 indicate that our proposed VAMP-EM detector success in accurately estimating the noise variance γ_n. Furthermore, the OTSM BER performance of both the single tap and of the AMP detectors is even worse than that of the classic OFDM system employing LMMSE detection at moderate SNR values for $v = 480$ km/h and at high SNR values at $v = 240$ km/h, respectively. This is because the BER performance of the single tap detector suffers from high ICI, and the AMP detector struggles to provide precise detection results since $\boldsymbol{H}$ is non-Gaussian. Finally, observe from Fig. 7 (a) and Fig. 7 (b) that the BER performance of detectors does not successively suffer upon increasing the vehicular velocity, which is consistent with the simulation results in [7], [10], [29].

Furthermore, in Fig. 8, we provide the BER performance of the LMMSE detector for fractional delay and Doppler shifts operating at different numbers of M and N, as well as different modulation orders. It is observed from Fig. 8 that given the values of M and N, a lower modulation order leads
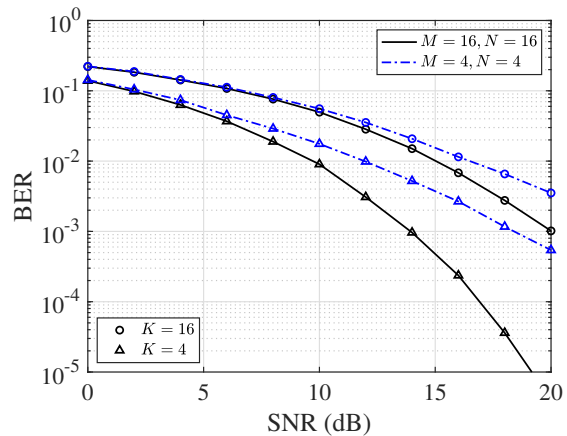

Fig. 8. BER performance of the LMMSE detector with non-integer delay and Doppler shifts operating at different modulation orders as well as different values of M and N.
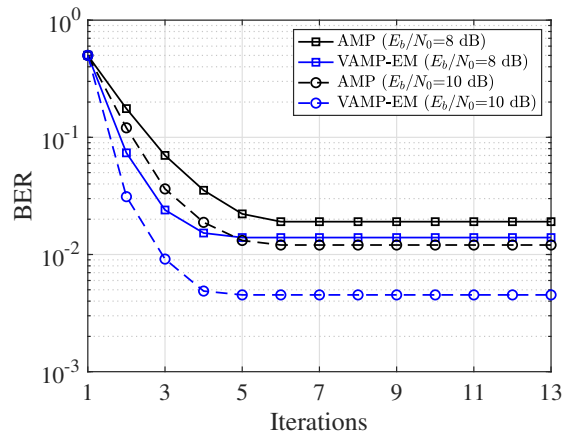

Fig. 9. BER performance of the AMP and the proposed VAMP-EM detector versus iteration number operation at $E_b/N_0=8$ dB and $E_b/N_0=9$ dB.

to a better BER performance, as expected. More specifically, at a BER of 10^{-3}, the $(M, N) = (16, 16)$ QPSK-modulated system is capable of achieving about 6 dB SNR gain compared to its 16QAM-modulated counterpart. Furthermore, the higher the value of MN, the better the BER performance becomes. Explicitly, under the condition of $K = 4$, the OTSM system using $(M, N) = (16, 16)$ requires about 4.5 dB lower SNR than the $(M, N) = (4, 4)$ system at a BER of 10^{-3}. This can be explained by the fact that a higher value of MN leads to a lower BER, as expected from (28).

In Fig. 9, we investigate the BER performance of both the AMP detector and of our proposed VAMP-EM detector versus

the number of detector iterations. The maximum number of iterations are $T_{\text{AMP}} = T = 13$, and the remaining parameters are the same as in Fig. 7 (a). Based on the results, we have the following observations. Firstly, the VAMP-EM detector achieves a better BER than the AMP detector, which is in accord with the observations from Fig. 7. Moreover, we can see that the proposed VAMP-EM detector converges for $T = 4$, yielding a better convergence performance than the AMP detector ($T_{\text{AMP}} = 6$). This is because we incorporate the auto-tuning part into the VAMP-EM detector, which allows the VAMP-EM to recover the transmitted symbols and to estimate the noise variance $1/\gamma_n$ jointly with an improved convergence rate.

Fig. 10 shows the detection complexity of LMMSE, VAMP-EM, AMP, GS, UAMP and single tap detectors. The parameters are consistent with those in Fig. 7 (a). It can be seen that the complexity of VAMP-EM is only slightly higher than that of the AMP, GS and UAMP detectors. But they are still of the same order. By contrast, the LMMSE detector imposes the highest complexity. This is because the VAMP-EM detector performs LMMSE and noise variance estimation at every iteration, while the UAMP detector does not involve noise variance estimation under this condition. Based on Fig. 7 and Fig. 10, it can be observed that the proposed VAMP-EM attains a better BER performance at the cost of a slighter higher complexity than that of the UAMP detector. On the other hand, since the complexity of our VAMP-EM detector is still higher than that of the single-tap detector, we will consider further reducing the VAMP-EM complexity in our future work. Moreover, based on Section IV-E, it can be readily observed that the complexities of the AMP and the VAMP-EM corresponding to large-scale OTSM systems are dominated by the value of MN^2L and M^2N^2, respectively. Based on Section II, it can be readily shown that the throughput of the 64QAM-modulated system is three times higher than that of 4QAM. Hence for the 64QAM OTSM system, the complexities per bit of the AMP-family detectors are nearly three times lower than that of 4QAM. Furthermore, based on Fig. 7, Fig. 9, and Fig. 10, we conclude that the VAMP-EM detector strikes a compelling BER vs. complexity trade-off.

Then, we evaluate the overall performance of the AMP and VAMP-EM turbo receivers in our rate-R LDPC-coded OTSM systems of Fig. 1 and the sum-product decoding algorithm is employed [30]. The remaining parameters are consistent with those in Fig. 7 (a). In Fig. 11, the EXIT curves of the proposed AMP-family-based detectors and that of the rate-1/2 LDPC decoder are investigated, where the superscripts *a* and *e* represent *a priori* and extrinsic information, while the subscripts *det* and *dec* indicate the detector and the decoder, respectively [31]. As shown in Fig. 11, the EXIT-tunnel between the detector's and the decoder's curve remain closed at E_b/N_0=2.1 dB. By contrast, an open EXIT-tunnel emerges at E_b/N_0=5.1 dB, which implies that the proposed AMP and VAMP-EM algorithms are capable of converging in this case. Moreover, in Fig. 11, the stair-case-shaped decoding trajectories at E_b/N_0=5.1 dB between the LDPC decoder and the inner detectors are provided for characterizing the processes of extrinsic information exchanges. It can be observed
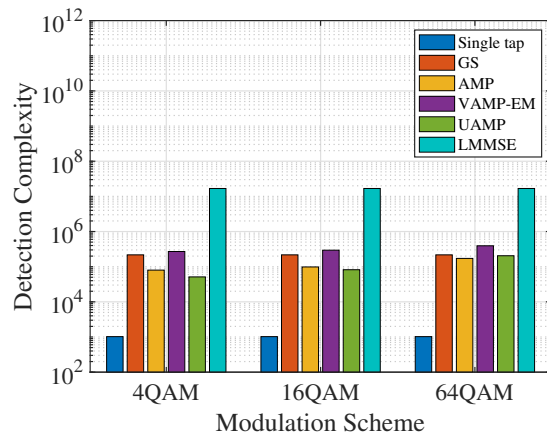

Fig. 10. Detection complexity of OTSM using LMMSE, single tap, GS, AMP, UAMP and VAMP-EM detectors.
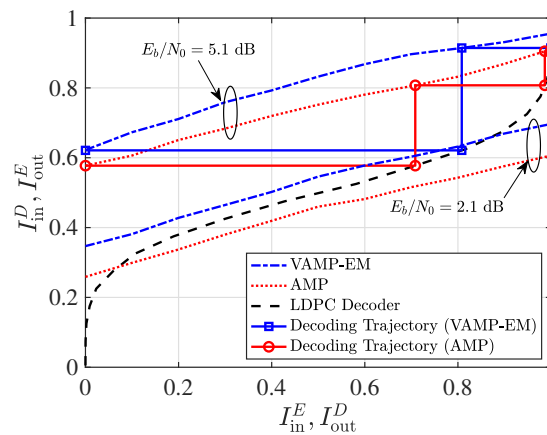

Fig. 11. EXIT charts of the proposed VAMP-EM detector and the AMP detector. The decoding trajectories are shown for the 1/2-rate LDPC decoder and for the proposed detector at $E_b/N_0 = 2.1$ dB and $E_b/N_0 = 5.1$ dB.

that the AMP turbo receiver requires at least five iterations between the detector and the LDPC decoder for approaching the maximum mutual information point, while only three corresponding iterations are required by the proposed VAMP-EM turbo receiver to converge around E_b/N_0=5.1 dB. This illustrates that the VAMP-EM detector is capable of attaining an improved convergence performance.

In Fig. 12, the BER performance of the rate-1/2 LDPC coded OTSM system using the proposed VAMP turbo receivers relying on different numbers of inner decoding iteration T_{in} and outer detection iteration T_{out} are evaluated. From Fig. 12, we have the following observations. Firstly, the BER performance of the VAMP-EM turbo receiver improves as T_{in} increases. Additionally, the BER performance can be significantly improved by using a higher value of T_{out}. However, the corresponding BER performance gain becomes marginal in the case of T_{out} being higher than 3, which implies that the combination of the number of inner iterations T_{in} and outer iterations T_{out} can be adjusted to strike a performance vs. complexity trade-off. Finally, there is no substantial system performance gain beyond E_b/N_0=5.1 dB, and all the above-mentioned observations are consistent with the predictions of the EXIT analysis of Fig. 11.

The BER performance of the AMP and VAMP turbo receivers using 1/2-rate and 3/4-rate LDPC coding is investi-
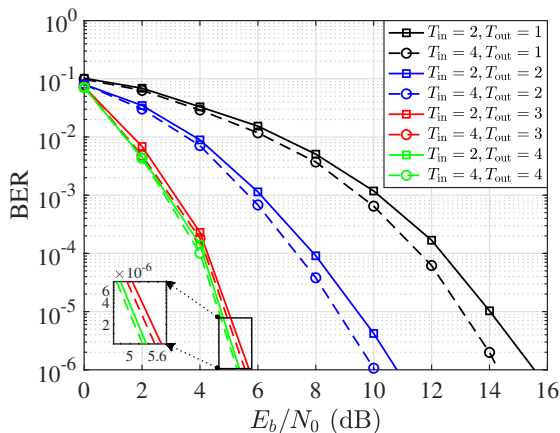
Fig. 12. BER performance of the proposed VAMP-EM detector for the rate-1/2 LDPC-coded OTSM system with different numbers of inner decoding iteration and outer turbo detection iteration.
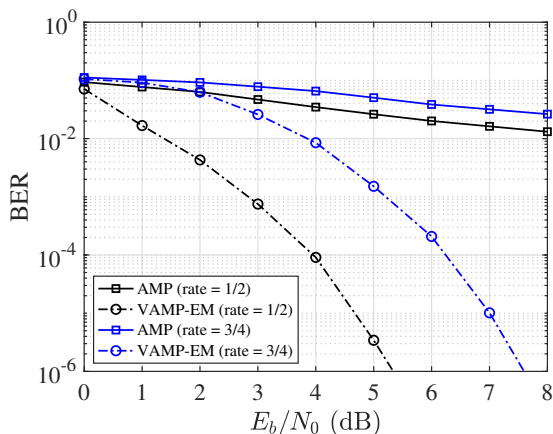

Fig. 13. BER performance of the rate-1/2 and rate-3/4 LDPC coded OTSM systems using the AMP detector and the proposed VAMP-EM detector.

gated in Fig. 13, where the numbers of inner and external iterations are set as $T_{\text{in}} = 4$ and $T_{\text{out}} = 4$, respectively. Based on Fig. 7 (a) and Fig. 13, we observe that the BER performance gaps between the AMP and VAMP-EM turbo receivers become larger than that in the uncoded system. Moreover, the lower the LDPC coded rate, the more significant the BER performance gap and the better the BER performance becomes. Furthermore, at a BER of 10^{-6}, it can be observed that the coding gain is about 2.2 dB associated with $R = 1/2$ compared to $R = 3/4$. Finally, similar to uncoded systems, due to the ill-conditioning of the DS-domain channel matrix $\boldsymbol{H}$, the convergence performance of the AMP turbo receiver cannot be ensured. Hence, we can observe an error floor at BER=10^{-2}.

VII. CONCLUSION

The performance analysis of OTSM systems was provided. Then, based on our theoretical derivations, the input-output relationship of generalized OTSM systems was established. Furthermore, the asymptotical BER upper bound of OTSM has been derived. Our simulation results have shown that the upper bound becomes increasingly tighter upon increasing the SNR. Then, an AMP-aided OTSM detection framework was conceived. By exploiting the statistics of the transmitted

symbols as the *a priori* information, a novel VAMP-EM detector was proposed. Explicitly, the EM algorithm was invoked for simultaneously recovering the transmitted symbol and for estimating the noise variance, making the proposed VAMP-EM detector more practical than its conventional counterparts. For further improving the convergence speed attained, an auto-tuning algorithm was conceived as part of the VAMP-EM approach. Our simulation results have demonstrated that the VAMP-EM algorithm is capable of attaining a better convergence and BER performance than the state-of-the-art detection schemes, despite its relatively low complexity. Then both an AMP and a VAMP-EM-based turbo receiver were proposed for LDPC-coded OTSM systems. Simulation results have been offered for characterizing the overall system performance. The VAMP-EM turbo receiver exhibits excellent convergence properties by exploiting the power of iterations between the detector and the decoder.

APPENDIX A
GENERALIZED EM FRAMEWORK BASED ON GIBBS SAMPLING

We consider solving the learning γ_n problem by leveraging the EM algorithm as illustrated in (51) and (52). Hence, the process of updating γ_n can be rewritten as

$$
\begin{aligned}
V\left(\gamma_n; \hat{\gamma}_n^t\right) &= -\mathbb{E}[\ln p(\boldsymbol{x}; \gamma_n)|\boldsymbol{y}; \hat{\gamma}_n^t] - \mathbb{E}[\ln p(\boldsymbol{y}|\boldsymbol{x}; \gamma_n)|\boldsymbol{y}; \hat{\gamma}_n^t] \\
&= -\mathbb{E}[\ln p(\boldsymbol{x}; \gamma_n)|q^t] - \mathbb{E}[\ln p(\boldsymbol{y}|\boldsymbol{x}; \gamma_n)|q^t] \\
&= J(q^t, q^t; \gamma_n) + \text{const}, \quad (66)
\end{aligned}
$$

where $q^t = p(\boldsymbol{x}|\boldsymbol{y}; \hat{\gamma}_n^t)$ and the Gibbs free energy is given by $J(q_1, q_2; \gamma_n) \triangleq D_{\text{KL}}[q_1 || p(\boldsymbol{y}|\boldsymbol{x}; \gamma_n)] + H(q_2)$. The EM processes of (51) and (52) can be expressed as [26]

$$
q^t = p(\boldsymbol{x}, \boldsymbol{y}; \gamma_n^t) \quad (67)
$$

$$
\hat{\gamma}_n^t = \arg\min_{\gamma_n} J(q^t, q^t; \gamma_n). \quad (68)
$$

It should be noticed that we can simplify $J(q_1, q_2; \gamma_n)$ as [27]

$$
\begin{aligned}
J(q^t, q^t; \gamma_n) &= -\ln p(\boldsymbol{y}; \gamma_n) + D_{\text{KL}}[q^t || p(\boldsymbol{y}|\boldsymbol{x}; \gamma_n)] \\
&\geq -\ln p(\boldsymbol{y}; \gamma_n) \quad (69)
\end{aligned}
$$

for any q^t since $D_{\text{KL}} \geq 0$. Therefore, a tighter upper bound of $-\ln p(\boldsymbol{y}; \gamma_n)$ is obtained by choosing q^t properly.

REFERENCES

[1] R. Steele and L. Hanzo, *Mobile radio communications: Second and third generation cellular and WATM systems: 2nd.* IEEE Press-John Wiley, 1999.

[2] L. Hanzo, B. Choi, T. Keller *et al.*, *OFDM and MC-CDMA for broadband multi-user communications, WLANs and broadcasting.* John Wiley & Sons, 2005.

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

[4] G. Surabhi, R. M. Augustine, and A. Chockalingam, "On the diversity of uncoded OTFS modulation in doubly-dispersive channels," *IEEE Transactions on Wireless Communications*, vol. 18, no. 6, pp. 3049–3063, 2019.

[5] F. Liu, Z. Yuan, Q. Guo, Z. Wang, and P. Sun, "Multi-block UAMP-based detection for OTFS with rectangular waveform," *IEEE Wireless Communications Letters*, vol. 11, no. 2, pp. 323–327, 2022.

[6] P. Raviteja, Y. Hong, E. Viterbo, and E. Biglieri, "Practical pulse-shaping waveforms for reduced-cyclic-prefix OTFS," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 1, pp. 957–961, 2019.

[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] 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 Transactions on Communications*, vol. 70, no. 1, pp. 635–648, 2022.

[9] 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.

[10] S. Li, J. Yuan, W. Yuan, Z. Wei, B. Bai, and D. W. K. Ng, "Performance analysis of coded OTFS systems over high-mobility channels," *IEEE Transactions on Wireless Communications*, vol. 20, no. 9, pp. 6033–6048, 2021.

[11] 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 Transactions on Wireless Communications*, 2021.

[12] S. Li, W. Yuan, Z. Wei, and J. Yuan, "Cross domain iterative detection for orthogonal time frequency space modulation," *IEEE Transactions on Wireless Communications*, vol. 21, no. 4, pp. 2227–2242, 2022.

[13] T. Thaj and E. Viterbo, "Orthogonal time sequency multiplexing modulation," in *2021 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2021, pp. 1–7.

[14] T. Thaj, E. Viterbo, and Y. Hong, "Orthogonal time sequency multiplexing modulation: Analysis and low-complexity receiver design," *IEEE Transactions on Wireless Communications*, vol. 20, no. 12, pp. 7842–7855, 2021.

[15] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[16] Z. Sui, S. Yan, H. Zhang, L.-L. Yang, and L. Hanzo, "Approximate message passing algorithms for low complexity OFDM-IM detection," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 9, pp. 9607–9612, 2021.

[17] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *IEEE Transactions on Information Theory*, vol. 65, no. 10, pp. 6664–6684, 2019.

[18] A. K. Fletcher, M. Sahraee-Ardakan, P. Schniter, and S. Rangan, "Rigorous dynamics and consistent estimation in arbitrarily conditioned linear systems," *CoRR*, vol. abs/1706.06054, 2017. [Online]. Available: http://arxiv.org/abs/1706.06054

[19] S. Sarkar, A. K. Fletcher, S. Rangan, and P. Schniter, "Bilinear recovery using adaptive Vector-AMP," *IEEE Transactions on Signal Processing*, vol. 67, no. 13, pp. 3383–3396, 2019.

[20] C. F. Van Loan, "The ubiquitous kronecker product," *Journal of Computational and Applied Mathematics*, vol. 123, no. 1-2, pp. 85–100, 2000.

[21] H. Zhang, L.-L. Yang, and L. Hanzo, "Compressed sensing improves the performance of subcarrier index-modulation-assisted OFDM," *IEEE Access*, vol. 4, pp. 7859–7873, 2016.

[22] H. Zhang, C. Jiang, L.-L. Yang, E. Basar, and L. Hanzo, "Linear precoded index modulation," *IEEE Transactions on Communications*, vol. 67, no. 1, pp. 350–363, 2018.

[23] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 744–765, 1998.

[24] H. Zhang, C. Jiang, J. Wang, L. Wang, Y. Ren, and L. Hanzo, "Multicast beamforming optimization in cloud-based heterogeneous terrestrial and satellite networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 1766–1776, 2020.

[25] 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.

[26] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in graphical models*. Springer, 1998, pp. 355–368.

[27] J. P. Vila and P. Schniter, "Expectation-maximization gaussian-mixture approximate message passing," *IEEE Transactions on Signal Processing*, vol. 61, no. 19, pp. 4658–4672, 2013.

[28] M. Tuchler, R. Koetter, and A. C. Singer, "Turbo equalization: principles and new results," *IEEE Transactions on Communications*, vol. 50, no. 5, pp. 754–767, 2002.

[29] T. Thaj and E. Viterbo, "Unitary-precoded single-carrier waveforms for high mobility: Detection and channel estimation," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, 2022, pp. 962–967.

[30] D. J. MacKay and R. M. Neal, "Near Shannon limit performance of low density parity check codes," *Electronics Letters*, vol. 33, no. 6, pp. 457–458, 1997.

[31] M. El-Hajjar and L. Hanzo, "EXIT charts for system design and analysis," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 127–153, 2013.

Zeping Sui (S'19) received the B. Eng. degree in electromagnetic field and wireless technology from Northwestern Polytechnical University, Xi'an, China, in 2017, and the Ph.D. degree in signal and information processing from University of Chinese Academy of Sciences, Beijing, China, in 2023. From March 2022 to Feb 2023, he was a research assistant with the Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), Singapore. His research interests include physical layer design of 5G and 6G wireless networks, emerging waveform-based systems, and cell-free massive MIMO.

Shefeng Yan (SM'10) received the B.Sc., M.Sc., and Ph.D. degrees all in electrical engineering from Northwestern Polytechnical University, China, in 1999, 2001, and 2005, respectively. He was a Post-doctoral Research Associate with the Institute of Acoustics, Chinese Academy of Sciences (IACAS), Beijing, China, from 2005 to 2007, and with the Department of Electronics and Telecommunications, Norwegian University of Science and Technology, Trondheim, Norway, from 2007 to 2009, respectively. He was a Senior Visiting Scholar with the Department of Electrical, Electronic and Communication Engineering, University of Erlangen-Nuremberg, Erlangen, Germany, from December 2015 to February 2016. Since 2009, he has been a Professor with IACAS, and with the University of Chinese Academy of Sciences. He has authored the book *Broadband Array Processing* (Springer, 2019). His research interests include acoustic signal processing, statistical and array signal processing, and underwater acoustic communications. Prof. Yan is a member of the Acoustical Society of America. He is a recipient of the 2010 ICA-ASA Young Scientist Grant for excellent contributions to Acoustics and a corecipient of the Best Paper Awards at SENSORCOMM 2008, WASPAA 2009, and WCSP 2018.

Hongming Zhang (S'15-M'17) received the B. Eng. degree (with first class honors) in telecommunications from the Nanjing University of Aeronautics and Astronautic (NUAA) and the City University London, both in 2011, the M. Sc. and Ph.D. degrees in wireless communications from University of Southampton in 2012 and 2017, respectively. From 2017 to 2018, he was with the Department of Electrical Engineering, Columbia University, New York, NY, USA. He is currently an As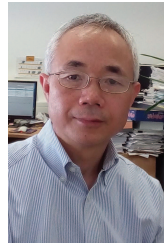sociate Professor with the School of Information and Communication Engineering. His research interests include maritime communications, heterogeneous networks, and machine learning.

Sumei Sun (F'16) is Distinguished Institute Fellow, and Acting Executive Director of Institute for Infocomm Research (I2R), Agency for Science, Technology, and Research (A*STAR), Singapore. Her current research interests are in next-generation wireless communications, joint communication-sensing-computing-control design, and industrial internet of things. She is an IEEE Fellow, Editor-in-Chief of IEEE Open Journal of Vehicular Technology, a member at large (MAL) with the IEEE Communications Society (2021-2023), and a member of the IEEE Vehicular Technology Society Board of Governors (2022-2024).

Yonghong Zeng (F'20) received the B.S. degree from Peking University, M.S. degree and Ph.D. degree from National University of Defense Technology, China. He worked in National University of Defense Technology, Nanyang Technological University and University of Hong Kong as assistant/associate professor or research follow before. He has been working in the Institute for Infocomm Research, A*STAR, since 2004, and currently is a senior scientist and group leader. He is working on B5G/6G communications, integrated sensing and communication, ultra-reliable low-latency communication, vehicular radar and communication, and real time localization system. He has co-authored six books and more than 90 refereed journal papers. He received the 2007 and 2009 IES (Institute of Engineers Singapore) prestigious engineering achievement award in Singapore, and ministry-level Scientific and Technological Development Awards four times in China. He received the award for advances in communication from IEEE Communication society in 2022, the Inaugural IEEE Communications Society Asia-Pacific Best Paper Award in 2012, and a few other best paper awards. He received the Certificate of Appreciation for outstanding contributions to the IEEE 802.22 standard in 2011. He is a Fellow of IEEE.

Lie-Liang Yang (M'98-SM'02-F'16) is the professor of Wireless Communications in the School of Electronics and Computer Science at the University of Southampton, UK. He received his MEng and PhD degrees in communications and electronics from Northern (Beijing) Jiaotong University, Beijing, China in 1991 and 1997, respectively, and his BEng degree in communications engineering from Shanghai TieDao University, Shanghai, China in 1988. He has research interest in wireless communications, wireless networks and signal processing for wireless communications, as well as molecular communications and nano-networks. He has published 400+ research papers in journals and conference proceedings, authored/co-authored three books and also published several book chapters. The details about his research publications can be found at https://www.ecs.soton.ac.uk/people/llyang. He is a fellow of both the IEEE and the IET, and was a distinguished lecturer of the IEEE VTS. He served as an associate editor to various journals, and is currently a senior editor to the IEEE Access and a subject editor to the Electronics Letters. He also acted different roles for organization of conferences.

Lajos Hanzo (Life Fellow, IEEE) received Honorary Doctorates from the Technical University of Budapest and Edinburgh University. He is a Foreign Member of the Hungarian Science-Academy, Fellow of the Royal Academy of Engineering (FREng), of the IET, of EURASIP and holds the IEEE Eric Sumner Technical Field Award.