Scalable Quadrature Spatial Modulation

Hyeon Seok Rou, *Graduate Student Member, IEEE*,
Giuseppe Thadeu Freitas de Abreu, *Senior Member, IEEE*, Hiroki Iimori, *Graduate Student Member, IEEE*,
David González G., *Senior Member, IEEE*, and Osvaldo Gonsa

***Abstract*— We consider quadrature spatial modulation (QSM) schemes, which achieve high spectral efficiency (SE) via the dispersion of a relatively small number P of M-ary modulated symbols over a large number of combinations of n_T transmit antennas and T transmit instances. In particular, we design a new space-time block code (STBC)-based scalable QSM scheme combining high SE with maximum diversity and optimum coding gains. Deriving a closed-form expression for the optimum SE, we show that scaling the size T with n_T not only is required to achieve SE optimality, but also results in further gains in bit error rate (BER) performance. Building on the latter optimal parameterization, a fully optimized scalable QSM (OS-QSM) transmitter design is then obtained by introducing a new dispersion matrix index selection algorithm that ensures even utilization of spatial-temporal resources. Finally, a new greedy boxed iterative shrinkage thresholding algorithm (GB-ISTA) QSM receiver is proposed, which exploits the inherent sparsity of QSM signals and while detecting spatially and digitally modulated bits in a greedy fashion. The resulting low complexity of the new receiver, which is linear on n_T, enables the utilization of OS-QSM in systems of previously prohibitive dimensions.**

***Index Terms*— Quadrature spatial modulation, space-time coding, massive MIMO, optimality, scalablility, low-complexity.**

I. Introduction

SPATIAL modulation (SM) [1]–[3] is a promising technique that can reduce the hardware complexity and costs in massive multiple-input multiple-output (MIMO) wireless communication systems without sacrificing BER and SE performances. In particular, in spatial modulation (SM) schemes, information bits are embedded not only in the selection of transmit symbols (a.k.a. constellation dimension), but also in the selection of transmit antennas utilized during transmission (a.k.a. spatial dimension).

Thanks to this approach, the vast spatial resources associated with massive MIMO systems can be efficiently utilized without requiring an equally large number of radio frequency (RF) chain components. This efficient utilization of RF-chains makes SM schemes attractive for future wireless systems such as beyond fifth generation (B5G), which will continue to make extensive use of millimeter-wave (mmWave) bands, and sixth generation (6G) networks [4], which are expected to also incorporate Terahertz and visible light communications (VLC) bands [5].

A major drawback of early SM schemes [1]–[3] is, however, that only one antenna is selected per transmission, which severely limits achievable SEs. In order to circumvent this limitation, a generalized spatial modulation (GSM) scheme was later developed [6], where combinations of multiple antennas are selected at each transmission, leading to substantial increase in SE. The notion was then developed into the QSM scheme [7], in which the real and imaginary components of the modulated symbols are allocated independently to different antennas via dedicated spatial-temporal dispersion matrices, resulting in a twofold increase in the encodable bits in the spatial (antenna) dimension, in addition to superior BER performance. This innovative QSM idea motivated further research, inspiring multiple clever successors including the Signed QSM [8] and the Extended Signed QSM [9], which increased the SE by exploiting positive and negative quadrature spaces; the Improved QSM [10] and the Fully Improved QSM [11], which aimed to increase the SE in a super-logarithmic proportion (previously logarithmic) with the number of transmit antennas; and the Full duplex (FD) QSM [12], which pioneered integration of QSM schemes to full duplex (FD) systems.

Despite various improvements contributed by the aforementioned works, these early QSM schemes still maintained an exclusive focus on increasing SE without a significant effort to reduce BER. Recognizing this limitation, the idea of combining SM/QSM with space-time codes (STCs) emerged [13], leading to schemes such as space-time shift keying (STSK) based on linear dispersion (LD) coding proposed in [14]–[16]; methods incorporating STBCs such as those in [17]–[20]; the spatial modulation with cyclic structure (CSM) of [21] and the Parallel QSM [22] scheme, where BERs are reduced by operating with simultaneously grouped subsets of the available transmit antennas.

The idea was also developed into QSM techniques with progressively enhanced dispersion matrix designs, which include the diversity-achieving quadrature spatial modulation (DA-QSM) scheme first proposed in [23] and later improved in [24] by the incorporation of Alamouti codes, and the more recent enhanced diversity-achieving quadrature spatial modu-

Manuscript received 24 July 2021; revised 17 December 2021 and 28 February 2022; accepted 6 May 2022. Date of publication 24 May 2022; date of current version 11 November 2022. The associate editor coordinating the review of this article and approving it for publication was M. Guillaud. *(Corresponding author: Hyeon Seok Rou.)*

Hyeon Seok Rou, Giuseppe Thadeu Freitas de Abreu, and Hiroki Iimori are with the School of Computer Science and Engineering, Jacobs University Bremen, 28759 Bremen, Germany (e-mail: h.rou@jacobs-university.de; g.abreu@jacobs-university.de; h.iimori@jacobs-university.de).

David González G. and Osvaldo Gonsa are with the Wireless Communications Technologies Group, Continental AG, 65936 Frankfurt am Main, Germany (e-mail: david.gonzalez.gonzalez@continental-corporation.com; osvaldo.gonsa@continental-corporation.com).

Digital Object Identifier 10.1109/TWC.2022.3175579

lation (EDA-QSM) method [25] in which dispersion matrices are constructed using the full-diversity full-rate (FDFR) codes with block-by-block sphere-decodability of [26].

It is important to emphasize that the aforementioned schemes focus exclusively on the transmit signal design, with the assumption of no channel state information (CSI) knowledge at the transmitter. There is, of course, another line of work on SM and QSM schemes in which transmit CSI is assumed and exploited to improved BER performance via signal shaping [27]. Similarly, there are also recent contributions in the design of QSM schemes suitable to the scenario of no receiver CSI [28].

In this article we focus on the more common scenario in which CSI is available only and perfectly at the receiver, for which the EDA-QSM can be considered as the state-of-the-art (SotA) *spatial-temporal modulation* scheme, as it was shown to outperform the aforementioned spatial-temporal modulation schemes both in BER and SE in the no transmit CSI case. But despite its advantages over earlier methods, the EDA-QSM and its predecessors still share two major shortcomings. The first is that the dispersion matrices used in QSM schemes proposed so far are based on 2×2 STBCs, which limits the achievable diversity and coding gains. With regards to that first limitation, we will in fact show in this article that QSM designs based on STBCs of a size T that do not scale with n_T are fundamentally *sub-optimal* in the SE sense.

The second shortcoming is that current QSM detection schemes are based either on exhaustive maximum likelihood (ML) or, at best, sphere-detectors. Here, it is worth noting that it has actually been proved in [29, Th.2], contrary to previous claims, that sphere decoding has an average complexity that still grows exponentially with the number of jointly decoded symbols. That result was corroborated by the findings of [30], where a cubic closed-form expression for the expected complexity of sphere detectors was derived, as well as of [31], where it was shown that lattice-reduction does not improve the tail exponent of the complexity distribution of sphere detectors. With regards to that second limitation, we will show in this article that in fact the complexity of ML- and sphere detection (SD)-based QSM receivers are both geometric on n_T and T, with P as exponent, such that these techniques are fundamentally *non-scalable* in the context of QSM systems. In other words, a severe and two-folded scalability challenge exists among current QSM schemes, namely, the absence of scalable transmitter and receiver designs.

Motivated by this challenge, we seek in this article to contribute a new QSM solution that is both, at the transmitter side, scalable to arbitrary system sizes ($i.e.$, with no limits on n_T, T and P), and, at the receiver side, decodable in polynomial time ($i.e.$, practical for large n_T and T, with moderate P). As a bonus, we also seek to furbish the proposed QSM scheme with every possibility to optimize SE, diversity and coding gains. To that end, we first introduce the optimal full-diversity full-rate (FDFR) Golden STBC of [32] in the design of the QSM dispersion matrices. The Golden code [32] is known to optimal ($i.e.$, FDFR with highest coding gain) over Gaussian constellations, and which was shown in [33], [34] to be constructible generally for arbitrary block sizes. To the best of our knowledge, the resulting OS-QSM scheme is the first method proposed so far which has this feature.

The new OS-QSM design is further enhanced with a new algorithm to select the indices of the dispersion matrices employed in the scheme, which ensures that all transmit antennas are utilized as often and with the same likelihood over the transmission of multiple blocks, thus ensuring optimally diverse utilization of all spatial-temporal resources. We again emphasize that such optimization is for the case of no transmit CSI, unlike that in [27] where dynamic optimization is done exploiting the available transmitter CSI to select the activated antenna indices.

Finally, in order to ensure feasible decodability matching the scalability of the transmitter design a new GB-ISTA QSM detector is proposed, which leverages on sparse recovery methods [35], [36]. Thanks to its sparse signal processing approach, the proposed decoding scheme does not rely on any restriction on the core code design, unlike preceding sphere-detection methods which requires block-diagonal fast-decodability. In addition, and most importantly, a major advantage of the proposed GB-ISTA receiver is that it does not require a search over the large codebook space, unlike the ML and state codewords matched block-by-block sphere decoding (SCMB-SD). In fact, the complexity order of the proposed receiver is shown to be cubic on T, quadratic on P, and linear on n_T. All in all, the contributions of the article can be summarized as follows:

- **Coding Gain Optimality**: In Subsection III-A, a transmission scheme is formulated following the suggestion given in [25] of substituting the Sezginer-Sari-Biglieri (SSB)-based core dispersion matrix with the 2×2 Golden code [32], which achieves optimal coding gain over integer symbol constellations. This design is enabled by the independence of the receiver later introduced in Section IV, which does not require such block-by-block orthogonality to function, as preceding methods did.
- **Scalability of Transmitter**: In Subsection III-B, the above coding-gain optimal design is generalized via the extension of the 2×2 Golden code into its $T \times T$ FDFR STBC variation [33], [34], making the design applicable to a wide range of parameters n_T, T and P.
- **Spectral Efficiency Optimality**: In Subsection III-C, a closed form expression for the optimum number of encoded symbols P^* required for a QSM to achieve SE optimality is given, which combined to the rate-optimality condition of STBCs, highlights the importance of systematic scalability of the STBC size T in the design of SE-optimal QSM schemes.
- **Resource Utilization Optimality**: In Algorithm 1, a new mechanism to select the optimal set of dispersion matrix indices is offered, ensuring that all Q spatial-temporal resources are equally utilized over time, for optimal transmit diversity gain without transmitter CSI.
- **Scalability of Receiver**: In Section IV, a new low-complexity greedy iterative shrinkage thresholding algorithm (ISTA)-based demodulation algorithm for QSM

schemes is proposed, which is feasible at large scales due to its low complexity. This decoding scheme can be applied to other STBC-QSM schemes with minor modifications.

- **Complexity of Receiver**: In Subsection IV-C, a novel complexity expression of the proposed receiver is derived and shown to be cubic on T, quadratic on P, and linear on n_T, in contrast to the ML and SD receivers which are geometric on T and n_T, with P as exponent.

Notation: Complex matrices and vectors are denoted in bold-face uppercase and lowercase letters, with their elements denoted by indexed normal lowercase letters, as in $\mathbf{X}$, $\mathbf{x}$ and x_i, respectively. The real and the imaginary parts of a complex number x are respectively denoted by x^R and x^I, respectively, and for the sake of future convenience we define for a complex vector $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T$ the associated decoupled vector $\boldsymbol{x} \triangleq [x_1^R, x_1^I, \cdots, x_n^R, x_n^I]^T$. The quadrature operator $\check{(\cdot)}$ applied to a complex number x yields the quadrature-separated form $\check{x}$, and will also be applied to $m \times n$ complex matrices $\mathbf{X}$, for which it yields the corresponding $2m \times 2n$ matrix $\check{\mathbf{X}}$ (see footnote[1]).

In turn, the complex conjugate, transpose, Hermitian, trace, vectorization, and the diagonalization operators are denoted by $(\cdot)^*$, $(\cdot)^{\mathsf{T}}$, $(\cdot)^{\mathsf{H}}$, $\mathrm{tr}(\cdot)$, $\mathrm{vec}(\cdot)$, and $\mathrm{diag}(\cdot)$, respectively, while the $n \times n$ identity and the $m \times n$-sized all-zero and all-one matrices are respectively denoted by $\mathbf{I}_n$, $\mathbf{0}_{m\times n}$, and $\mathbf{1}_{m\times n}$. The p-norm with $p \geq 0$ is denoted by $||\cdot||_p$, while $|\cdot|$ denotes either element-wise absolute value operation (for vectors) or cardinality (for sets), respectively, and the sets of real, complex, and integer numbers are denoted by $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{Z}$, respectively. Expectation is denoted as $\mathbb{E}[\cdot]$, the floor to the nearest power of 2 is represented by $\lfloor\cdot\rfloor_{2\times}$, the conversion operation of a left-most-significant binary vector to the corresponding base-10 integer is denoted by $[\cdot]_{(10)}$. The binomial coefficient is denoted by $\binom{Q}{P}$, and $\otimes$ denotes the Kronecker product. The projection of a scalar v onto the set $\mathcal{X}$ is denoted by $\mathcal{P}_{\mathcal{X}}(v)$, and the complex Gaussian distribution with mean μ and variance σ^2 is denoted by $\sim \mathcal{CN}(\mu, \sigma^2)$.

II. Preliminaries

A. System Model

Consider a point-to-point (P2P) MIMO communication system in which a transmitter equipped with n_T transmit antennas exchanges information with a receiver equipped with n_R receive antennas employing SM. The received signal corresponding to T consecutive time slots (transmission instances) during which the channel is assumed to be constant can be compactly written as

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{V} \in \mathbb{C}^{n_R \times T}, \tag{1}$$

where $\mathbf{Y} \in \mathbb{C}^{n_R\times T}$ is the matrix collecting the signals received at each antenna and time slot, $\mathbf{H} \in \mathbb{C}^{n_R\times n_T}$ is the flat-fading channel matrix with elements $h_{i,j} \sim \mathcal{CN}(0,1)$, $\mathbf{X} \in \mathbb{C}^{n_T\times T}$ is the space-time transmit signal, and

[1]To be clear, $\check{x} \triangleq \begin{bmatrix} x^R & -x^I \\ x^I & x^R \end{bmatrix}$ and $\check{\mathbf{X}} \triangleq \begin{bmatrix} \check{x}_{1,1} & \cdots & \check{x}_{1,n} \\ \vdots & \ddots & \vdots \\ \check{x}_{m,1} & \cdots & \check{x}_{m,n} \end{bmatrix}$.

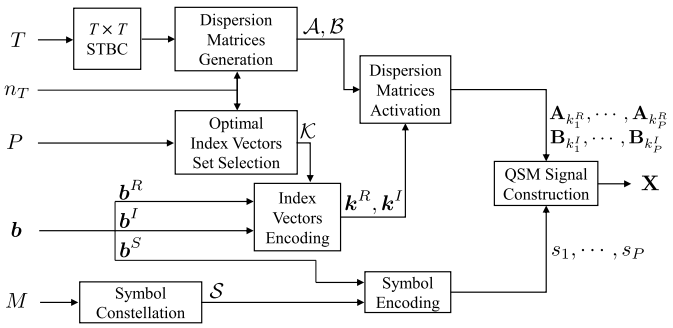


Fig. 1. Schematic diagram depicting the generic structure of a QSM transmission scheme.

$\mathbf{V} \in \mathbb{C}^{n_R\times T}$ is the additive white Gaussian noise (AWGN) with elements $v_{i,j} \sim \mathcal{CN}(0, N_0)$, where N_0 is the noise variance.

It is assumed hereafter that the quasi-static Rayleigh fading channel matrix $\mathbf{H}$ is known at the receiver but not at the transmitter, and we remark that since the channel power per matrix entry is unitary, the fundamental signal-to-noise ratio (SNR) is given by $\rho \triangleq \frac{1}{N_0}\mathbb{E}\left[\mathrm{tr}(\mathbf{X}^{\mathsf{H}}\mathbf{X})\right]$. In turn, in accordance with related QSM literature [7], [18]–[21], [23], [25] and as illustrated in Fig. 1, the transmit signal matrix $\mathbf{X}$ is constructed in a manner to convey the information of a bit sequence $\boldsymbol{b}$, both in the form of P digitally modulated signals, as well as in the form of the dispersed allocations to different antennas and time slots, as described by

$$\mathbf{X} = \sum_{p=1}^{P} (s_p^R \mathbf{A}_{k_p^R} + s_p^I \mathbf{B}_{k_p^I}), \tag{2}$$

where $s_p = s_p^R + j s_p^I$, with $p = \{1, \cdots, P\}$, are transmit symbols chosen from a complex constellation $\mathcal{S}$ of cardinality $|\mathcal{S}| = M$; $\mathbf{A}_{k_p^R}$ and $\mathbf{B}_{k_p^I}$ are dispersion matrices belonging to the sets $\mathcal{A} = \{\mathbf{A}_q\}_{q=1}^{Q} \in \mathbb{C}^{n_T\times T}$ and $\mathcal{B} = \{\mathbf{B}_q\}_{q=1}^{Q} \in \mathbb{C}^{n_T\times T}$, with $Q \triangleq T \times n_T$; and the indices k_p^R and k_p^I are the p-th elements of the index vectors $\boldsymbol{k}^R$ and $\boldsymbol{k}^I$, respectively, which are selected from an optimized set of index vectors $\mathcal{K} = \{\boldsymbol{k}_n\}_{n=1}^{N} \in \mathbb{N}^P$, with $N \triangleq \lfloor\binom{Q}{P}\rfloor_{2\times}$.

With regards to (2), and again referring to Fig. 1, we clarify that in QSM schemes the bit sequence $\boldsymbol{b}$ is subdivided into three sequences: $\boldsymbol{b}^S$, of length $B^S \triangleq P\log_2|\mathcal{S}| = P\log_2 M$, which corresponds to the information encoded in the symbols $\mathbf{s} = \{s_1, \cdots, s_P\}$, taken from $\mathcal{S}$, and the conjugate sequences $\boldsymbol{b}^R$ and $\boldsymbol{b}^I$, both of length $B^{\mathcal{K}} \triangleq \log_2|\mathcal{K}| = \log_2 N$, which correspond to the information encoded in the selection of spatial-temporal resources according to the dispersion matrix index vectors $\boldsymbol{k}^R$ and $\boldsymbol{k}^I$ from $\mathcal{K}$, respectively.

In view of the above, it can be said that the design of a specific QSM scheme amounts essentially to the method employed in the construction of each of the Q dispersion matrices $\mathbf{A}_q$ and $\mathbf{B}_q$ in the sets $\mathcal{A}$ and $\mathcal{B}$, and the selection of the set $\mathcal{K}$ containing the index vectors $\boldsymbol{k}^R$ and $\boldsymbol{k}^I$ which inform the choices of dispersion matrices used in each transmission.

B. State of the Art Review

To exemplify how SotA QSM schemes can be cast into the general framework described by (2), first consider the original QSM scheme proposed in [7]. In this case, the dispersion matrices reduce to dispersion vectors ($i.e., T = 1$ with $Q = n_T$) which are given by

$$\mathbf{A}_q = \mathbf{e}_q \quad \text{and} \quad \mathbf{B}_q = j\mathbf{e}_q, \tag{3}$$

where $\mathbf{e}_q$ is the q-th column of $\mathbf{I}_Q$, and no specific design criteria are given for the selection of the indices in the index vectors $\boldsymbol{k}^R$ and $\boldsymbol{k}^I$.

In turn, in the DA-QSM scheme of [23], two-column dispersion matrices ($i.e., T = 2$ with $Q = n_T$) are employed so as to exploit transmit diversity. In particular, in this scheme

$$\begin{aligned} \mathbf{A}_q &= \mathbf{M}_{n_T}^{q-1}\tilde{\mathbf{A}} \quad \text{and} \\ \mathbf{B}_q &= j\mathbf{M}_{n_T}^{q-1}\tilde{\mathbf{B}}, \end{aligned} \tag{4}$$

with $\tilde{\mathbf{A}} \triangleq \left[\begin{smallmatrix} \mathbf{I}_2 \\ \mathbf{0}_{(n_T-2)\times 2} \end{smallmatrix}\right]$ and $\tilde{\mathbf{B}} \triangleq \left[\begin{smallmatrix} \mathbf{M}_2 \\ \mathbf{0}_{(n_T-2)\times 2} \end{smallmatrix}\right]$, where $\mathbf{M}_n$ is an $n \times n$ cyclic lower-shift matrix,[2] such that its own $(q-1)$-th power pre-multiplied to a given matrix results in a shift of the bottom $(q-1)$ rows of the latter to the top.

From the above it is visible that the DA-QSM scheme of [23] improves over the QSM scheme of [7] essentially by adding diversity, $i.e.$, by extending the transmission instances from $T = 1$ to $T = 2$. However, the dispersion matrices of the DA-QSM method are still real, just as those of the QSM scheme of [7], implying that no additional multiplexing capability is aggregated, and that the coding gain is not optimized.

In contrast, the EDA-QSM method of [25] improves over the latter on both aspects. In particular, in this scheme the dispersion matrices are more elaborately designed as

$$\mathbf{A}_q = \mathbf{A}_{4(\ell-1)+i} = \mathbf{e}_\ell \otimes \mathbf{C}_i, \tag{5a}$$

$$\mathbf{B}_q = \mathbf{B}_{4(\ell-1)+i} = \mathbf{e}_\ell \otimes \mathbf{D}_i, \tag{5b}$$

where $\mathbf{e}_\ell$ is the ℓ-th column of $\mathbf{I}_L$, with $L \triangleq n_T/2$, the indices $i \in \{1, \cdots, 4\}$ and $\ell \in \{1, \cdots, L\}$, and the core matrices $\mathbf{C}_i$ and $\mathbf{D}_i$ are based on the SSB STBC of [26] described by

$$\mathbf{S}_{SSB} = \begin{bmatrix} s_1 + bs_3 & -s_2^* + jbs_4^* \\ s_2 + bs_4 & s_1^* - jbs_3^* \end{bmatrix} = \sum_{i=1}^{4} s_i^R \mathbf{C}_i + s_i^I \mathbf{D}_i, \tag{6}$$

with $b \triangleq \frac{(1-\sqrt{7})+j(1+\sqrt{7})}{4}$, and

$$\begin{aligned} &\mathbf{C}_1 \triangleq \mathbf{I}_2, \quad \mathbf{C}_2 \triangleq \mathbf{Z}, \\ &\mathbf{C}_3 \triangleq b\mathbf{W}, \quad \mathbf{C}_4 \triangleq \mathbf{Z} \cdot \mathbf{C}_3, \end{aligned} \tag{7a}$$

and

$$\begin{aligned} &\mathbf{D}_1 \triangleq j\,\mathbf{M}_2 \cdot \mathbf{Z}, \quad \mathbf{D}_2 \triangleq \mathbf{Z} \cdot \mathbf{D}_1, \\ &\mathbf{D}_3 \triangleq -b\mathbf{M}_2 \cdot \mathbf{W} \cdot \mathbf{M}_2, \quad \mathbf{D}_4 \triangleq \mathbf{Z} \cdot \mathbf{D}_3, \end{aligned} \tag{7b}$$

where $\mathbf{Z} \triangleq \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ and $\mathbf{W} \triangleq \begin{bmatrix} 1 & 0 \\ 0 & -j \end{bmatrix}$.

[2]To be clear, $\mathbf{M}_n$ is obtained by circularly shifting the bottom row of $\mathbf{I}_n$ to the top, such that, $e.g.$, $\mathbf{M}_3 = \left[\begin{smallmatrix} 0\,0\,1 \\ 1\,0\,0 \\ 0\,1\,0 \end{smallmatrix}\right]$.

Thanks to the concise description above it becomes easy to see that the fundamental distinction between the DA-QSM and the EDA-QSM methods is that the dispersion matrices of EDA-QSM are complex-valued, such that the orthogonality between the real and imaginary dimensions are better exploited in order to reap multiplexing and coding gains.

Two fair criticisms that can be made of the aforementioned schemes – and in fact, to the best of our knowledge of all existing SotA QSM methods proposed so far – are, however: a) that the scheme does not scale systematically simultaneously over space and time, for arbitrary $T > 2$; and b) that the coding gain achieved is not optimum. Mitigating these two limitations are the first two objectives of our first contribution described in the following section.

III. Optimized Scalable Quadrature Spatial Modulation (OS-QSM): Motivation and Transmitter Design

A. Coding Gain Optimal Spatial-Temporal QSM ($T = 2$)

The main aim of this subsection is to introduce the notion of *coding-gain optimality* to the SotA ($T \leq 2$) QSM schemes. To this end, let us begin by formulating an intermediate design, which improves upon the SotA EDA-QSM scheme by replacing the SSB code utilized in the latter with the coding gain optimal 2×2 Golden code [32], as conjectured[3] in [25].

The Golden code compactly encodes four symbols $\{s_1, s_2, s_3, s_4\}$ into the matrix,

$$\mathbf{S}_G = \frac{1}{\sqrt{5}} \begin{bmatrix} \alpha(s_1 + s_2\theta) & \alpha(s_3 + s_4\theta) \\ j\bar{\alpha}(s_3 + s_4\bar{\theta}) & \bar{\alpha}(s_1 + s_2\bar{\theta}) \end{bmatrix}, \tag{8}$$

where θ and $\bar{\theta}$ denote the complementary Golden numbers $\theta = (1+\sqrt{5})/2$ and $\bar{\theta} = (1-\sqrt{5})/2$, respectively, and $\alpha = 1 + j(1-\theta)$ and $\bar{\alpha} = 1 + j(1-\bar{\theta})$ are the optimized coefficients for the Gaussian integer constellation sets.

We highlight that the Golden code described in (8) is known to have maximum achievable coding gain among all 2×2 FDFR STBCs under integer Gaussian constellations, hence outperforming the SSB code utilized in the EDA-QSM.

The construction of QSM dispersion matrices based on the Golden code follows from the decomposition of $\mathbf{S}_G$ into the auxiliary matrices $\mathbf{C}_i$ and $\mathbf{D}_i$, which are used to modulate the real part s_i^R and imaginary part s_i^I of each i-th symbol encoded, respectively, such that

$$\mathbf{S}_G = \frac{1}{\sqrt{5}} \sum_{i=1}^{4} (s_i^R \mathbf{C}_i + s_i^I \mathbf{D}_i), \tag{9}$$

where

$$\begin{aligned} &\mathbf{C}_1 \triangleq \begin{bmatrix} \alpha & 0 \\ 0 & \bar{\alpha} \end{bmatrix}, \quad \mathbf{C}_2 \triangleq \boldsymbol{\Theta} \cdot \mathbf{C}_1, \\ &\mathbf{C}_3 \triangleq \mathbf{J}_{2,1} \cdot \mathbf{C}_1 \cdot \mathbf{M}_2, \quad \mathbf{C}_4 \triangleq \mathbf{J}_{2,1} \cdot \mathbf{C}_2 \cdot \mathbf{M}_2, \end{aligned} \tag{10a}$$

and

$$\mathbf{D}_i \triangleq j\mathbf{C}_i, \tag{10b}$$

[3]In the concluding remarks of [25] it was conjectured, without any further details or related development, that the EDA-QSM scheme proposed thereby could be improved via the utilization of other STBCs. The contribution of this subsection concretely realizes that suggestion, and is an intermediate step to the more significant generalization for $T > 2$ that follows hereafter.

with $\mathbf{\Theta} \triangleq \begin{bmatrix}\theta & 0\\ 0 & \bar{\theta}\end{bmatrix}$ and $\mathbf{J}_{2,1} \triangleq \begin{bmatrix}1 & 0\\ 0 & j\end{bmatrix}$. Note that post-multiplying the circular lower-shift matrix $\mathbf{M}_n$ to a given matrix $\mathbf{X}$ results in a columnwise left-circular shift of $\mathbf{X}$.

In possession of the above auxiliary matrices, the Golden dispersion matrices are then built using the Kronecker product operations following a similar strategy of [24], [25], namely

$$\mathbf{A}_q = \mathbf{A}_{4(\ell-1)+i} = \sqrt{\frac{2}{5}}\,\mathbf{e}_\ell \otimes \mathbf{C}_i \tag{11a}$$

$$\mathbf{B}_q = \mathbf{B}_{4(\ell-1)+i} = \sqrt{\frac{2}{5}}\,\mathbf{e}_\ell \otimes \mathbf{D}_i. \tag{11b}$$

Regarding the scaling factor in (11), the denominator $\frac{1}{\sqrt{5}}$ is passed over from the coefficient of the Golden code as in (8) and (9), while the numerator $\sqrt{2}$ is the result of power scaling required to ensure that the transmit power constraint $\mathbb{E}\left[\mathrm{tr}(\mathbf{X}^{\mathsf{H}}\mathbf{X})\right] = PT$ is satisfied. To elaborate further, (2) and (11) following $\mathbb{E}\left[\mathrm{tr}(\mathbf{X}^{\mathsf{H}}\mathbf{X})\right] = PT$ imply that the dispersion matrices must satisfy $\mathrm{tr}(\mathbf{A}_q^{\mathsf{H}}\mathbf{A}_q) = T$ and $\mathrm{tr}(\mathbf{B}_q^{\mathsf{H}}\mathbf{B}_q) = T$, for all $q \in \{1, \cdots, Q\}$, whereas from the construction of the auxiliary matrices $\mathbf{C}_i$ and $\mathbf{D}_i$ as per (10) it is evident that $\mathrm{tr}(\mathbf{C}_i^{\mathsf{H}}\mathbf{C}_i) = 1$ and $\mathrm{tr}(\mathbf{D}_i^{\mathsf{H}}\mathbf{D}_i) = 1$, such that a power scaling of $T = 2$ onto $\mathbf{C}_i$ and $\mathbf{D}_i$, *i.e.*, an amplitude scaling of $\sqrt{2}$ onto $\mathbf{C}_i$ and $\mathbf{D}_i$, is needed.

We may summarize the contribution of this subsection as an extension of the EDA-QSM scheme via the substitution of its core dispersion matrix with a Golden code-based alternative, alluded to in [25], that lends it coding gain optimality. Although it is trivial that this scheme, which we refer to as the Golden code-based quadrature spatial modulation (G-QSM), outperforms the original EDA-QSM, a direct simulated comparison between the two is provided in Fig. 11, found in Section III.

B. Scalable Coding Gain Optimal QSM Design ($T \geq 2$)

Despite the coding gain optimality it introduces as a STBC-based QSM scheme, the G-QSM design formulated in Subsection III-A still suffers from a limitation in the temporal dimension T, *i.e.*, the number of transmission instances, which is limited to $T = 2$. In this subsection, we therefore extend the notion of coding gain optimal QSM, generalized to scaled temporal dimensions, *i.e.*, $T \geq 2$. Straightforwardly, this restriction on T can be eliminated by designing the QSM dispersion matrices based on the *Perfect* FDFR STBC proposed in [33], [34]. Leaving details of the code construction to [33], [34], a $T \times T$ FDFR STBC encodes T^2 symbols such that the average energy transmitted per antenna is normalized to unity, an energy efficiency-shaping constraint is enforced, and a SE-preserving lower bound on the coding gain (a.k.a, non-vanishing determinant) is maximized. We highlight the last property which is the coding gain optimality for a given T, hereafter considered always to be a divisor of n_T.

Ultimately, for given $T \in \mathbb{N}^+$ the design is described by

$$\mathbf{S}_P = \sum_{t=1}^{T} \mathrm{diag}(\mathbf{R}\cdot\mathbf{s}_t)\cdot\mathbf{J}_{T,t-1}\cdot\mathbf{N}_T^{t-1}, \tag{12}$$

where $\mathbf{s}_t = [s_{1+(t-1)T}, s_{2+(t-1)T}, \cdots, s_{tT}]^{\mathsf{T}}$, with $t = \{1, \cdots, T\}$ are vectors each carrying T distinct transmit symbols, $\mathbf{R}$ is a $T\times T$ optimum lattice generating matrix [33], $\mathbf{J}_{T,n}$ is a $T \times T$ matrix constructed by replacing the last n diagonal entries of the identity matrix by the elementary complex number j, and $\mathbf{N}_T$ is a $T \times T$ cyclic upper-shift matrix,[4]such that post-multiplying it to a given matrix $\mathbf{X}$ results in a column-wise shift of $\mathbf{X}$ to the right.

Notice that the Perfect FDFR STBC of [33] fully generalizes the 2×2 Golden code of [32]. To see that, suffice it to consider the case $T = 2$ and with the corresponding lattice generating matrix $\mathbf{R} \triangleq \frac{1}{\sqrt{5}}\begin{bmatrix}\alpha & \alpha\theta\\ \bar{\alpha} & \bar{\alpha}\bar{\theta}\end{bmatrix}$, such that (12) yields

$$\begin{aligned}\mathbf{S}_P = &\ \mathrm{diag}\left(\frac{1}{\sqrt{5}}\begin{bmatrix}\alpha & \alpha\theta\\ \bar{\alpha} & \bar{\alpha}\bar{\theta}\end{bmatrix}\begin{bmatrix}s_1\\ s_2\end{bmatrix}\right)\cdot\mathbf{J}_{2,0}\cdot\mathbf{N}_2^0\\ &+\mathrm{diag}\left(\frac{1}{\sqrt{5}}\begin{bmatrix}\alpha & \alpha\theta\\ \bar{\alpha} & \bar{\alpha}\bar{\theta}\end{bmatrix}\begin{bmatrix}s_3\\ s_4\end{bmatrix}\right)\cdot\mathbf{J}_{2,1}\cdot\mathbf{N}_2^1\\ = &\ \frac{1}{\sqrt{5}}\begin{bmatrix}\alpha(s_1+s_2\theta) & \alpha(s_3+s_4\theta)\\ j\bar{\alpha}(s_3+s_4\bar{\theta}) & \bar{\alpha}(s_1+s_2\bar{\theta})\end{bmatrix} = \mathbf{S}_G.\end{aligned} \tag{13}$$

It follows that in order to be employ Perfect FDFR STBCs in the design of QSM, suffice it to decompose the core code structure of (12) in terms of corresponding auxiliary dispersion matrices $\mathbf{C}_i$ and $\mathbf{D}_i$ due to symmetry, namely

$$\mathbf{C}_i = \mathbf{C}_{T(w-1)+t} = \mathrm{diag}(\mathbf{R}\cdot\mathbf{e}_t)\cdot\mathbf{J}_{T,w-1},\cdot\mathbf{N}_T^{w-1}, \tag{14a}$$

$$\mathbf{D}_i = j\mathbf{C}_i, \tag{14b}$$

where the generalized indices $i \in \{1, \cdots, T^2\}$ are constructed systematically on $t \in \{1, \cdots, T\}$ and $w \in \{1, \cdots, T\}$, and $\mathbf{e}_t$ is the t-th column of $\mathbf{I}_T$.

Following this, the full set of dispersion matrices $\mathcal{A}$ and $\mathcal{B}$ can be built, *i.e.*,

$$\mathbf{A}_q = \mathbf{A}_{T^2(\ell-1)+i} = \gamma\mathbf{e}_\ell \otimes \mathbf{C}_i, \tag{15a}$$

$$\mathbf{B}_q = \mathbf{B}_{T^2(\ell-1)+i} = \gamma\mathbf{e}_\ell \otimes \mathbf{D}_i, \tag{15b}$$

where again $q \in \{1, \cdots, Q\}$, $i \in \{1, \cdots, T^2\}$ as from (14), $\mathbf{e}_\ell$ is the ℓ-th column of $\mathbf{I}_L$, but $\ell \in \{1, \cdots, L\}$ with $L = n_T/T$, as well as a generalized scaling factor γ determined depending on the specific STBC in order to adjust the powers of the dispersion matrices such that $\mathrm{tr}(\mathbf{A}_q^{\mathsf{H}}\mathbf{A}_q) = T$ and $\mathrm{tr}(\mathbf{B}_q^{\mathsf{H}}\mathbf{B}_q) = T$.

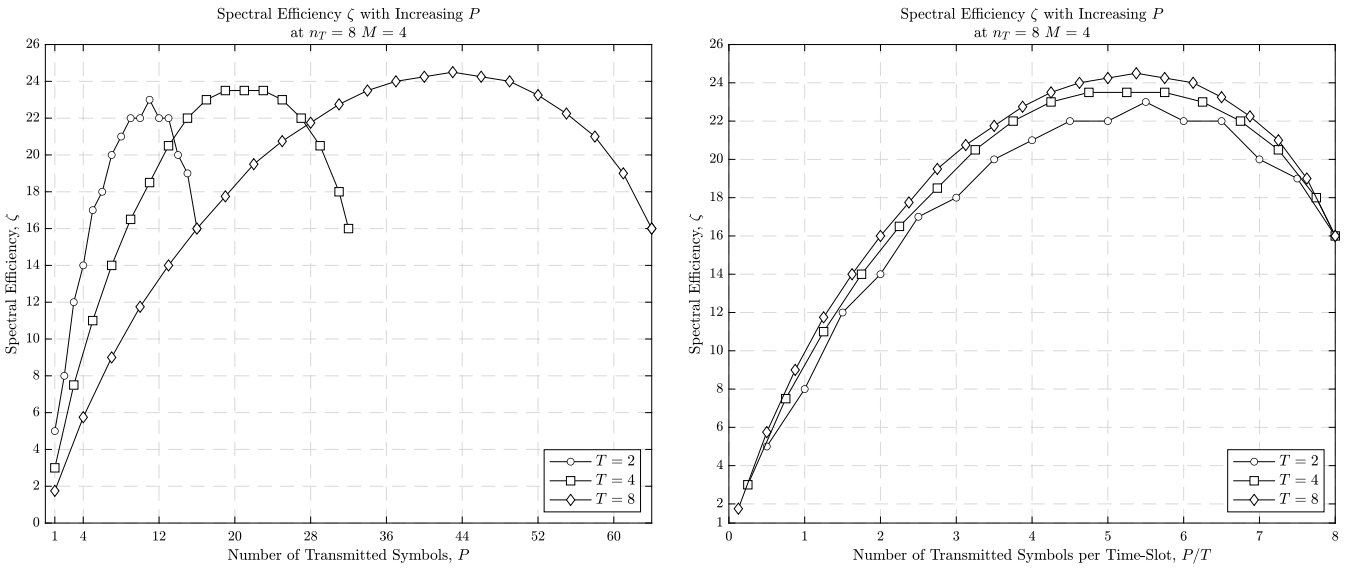
C. Spectral Efficiency Optimality of QSM Schemes

We now turn to the spectral efficiency optimality of QSM schemes and analyze the effects of the scalable parameters n_T, T and P, deriving corresponding SE-optimal parametrization.

Recall the QSM transmit symbol $\mathbf{X}$ as per (2), which conveys a total of B bits. With the fact that such transmission requires T successive channel uses, the SE ζ of any QSM scheme, in bits/s/Hz, is given by

$$\zeta(P, T; M, n_T) = \frac{B}{T} = \frac{1}{T}\left(2\left\lfloor\log_2\binom{Q}{P}\right\rfloor + P\log_2 M\right), \tag{16}$$

[4]Notice that $\mathbf{J}_{T,n}$ generalizes $\mathbf{J}_{2,1}$ used in (10). In turn, oppositely to $\mathbf{M}_n$, $\mathbf{N}_n$ is obtained by circularly shifting the top row of $\mathbf{I}_n$ to the bottom. Some examples are $\mathbf{J}_{2,0} \triangleq \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$, $\mathbf{J}_{3,2} \triangleq \begin{bmatrix}1 & 0 & 0\\ 0 & j & 0\\ 0 & 0 & j\end{bmatrix}$, and $\mathbf{N}_3 \triangleq \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 1 & 0 & 0\end{bmatrix}$.



(a) ζ versus P. (b) ζ versus P/T.

Fig. 2. Spectral efficiency of OS-QSM scheme with $T = 2,\ 4$, and 8, for a given system with $n_T = 8$ and $M = 4$.

where we recall that $Q \triangleq Tn_T$ and adopt a notation meant to emphasize that P and T are seen as fundamental QSM design parameters, while M and n_T are system constraints.

The presence of the binomial coefficient $\binom{Q}{P}$ in (16) implies that the SE function $\zeta(P, T; M, n_T)$ is monotonically descending on T for a fixed P, and concave on P for a fixed T, as well as on the ratio P/T. This is well illustrated in the plots offered in Fig. 2, from which it can be seen that in a system with $n_T = 8$ and $M = 4$, the highest attainable SEs, denoted by ζ^*, are achieved with $P = 11, 21$ and 42, for $T = 2, 4$ and 8, respectively.

The increasing effect onto SE by growing T, for a given system of n_T transmit antennas, even under constant P/T, is also visible in Fig. 2.

Motivated by the discussion above, we seek analytical expressions for the optimum ratio P/T that maximizes the SE, given n_T and M, which in turn can be used to determine the relative SE reduction incurred in setting $T < n_T$ for large-scale systems with $n_T \to \infty$. To this end, consider the upper and lower-bounds on the binomial coefficient discovered in [37], namely

$$\begin{aligned}\tfrac{e^{-1/8}}{\sqrt{2\pi P}}\Big(\tfrac{Q}{Q-P}\Big)^{Q+\frac{1}{2}}\cdot\Big(\tfrac{Q-P}{P}\Big)^{P} &< \binom{Q}{P} \\ &< \underbrace{\tfrac{1}{\sqrt{2\pi P}}\Big(\tfrac{Q}{Q-P}\Big)^{Q+\frac{1}{2}}\cdot\Big(\tfrac{Q-P}{P}\Big)^{P}}_{\triangleq\beta(P;Q)},\ \ \forall\, 1 \le P < Q,\end{aligned} \tag{17}$$

where, for future convenience we implicitly defined the upper-bounding function $\beta(P; Q)$.

Using (17) into (16) yields the bound

$$\zeta(P, T; M, n_T) < \overbrace{\frac{1}{T}\log_2\left(\beta^2(P; Q)\cdot M^P\right)}^{\triangleq\zeta^+(P,T;M,n_T)}, \tag{18}$$

and taking the derivative of the latter expression with respect to P yields

$$\begin{aligned}\frac{\partial\zeta^+}{\partial P} &= \frac{1}{T}\cdot\frac{\partial}{\partial P}\Big[2\log_2\Big(\tfrac{1}{\sqrt{2\pi P}}\big(\tfrac{Q}{Q-P}\big)^{Q+\frac{1}{2}}\big(\tfrac{Q-P}{P}\big)^{P}\Big) \\ &\qquad + P\log_2(M)\Big] \\ &= \frac{1}{T\ln(2)}\Big[2\Big(\ln\big(\tfrac{1-\varepsilon}{\varepsilon}\big) + \tfrac{1-2\varepsilon}{2Q(\varepsilon-1)\varepsilon}\Big) + \ln(M)\Big] \\ &= \frac{1}{T\ln(2)}\Big[\ln(\tfrac{(1-\varepsilon)^2\cdot M}{\varepsilon^2}) + \big(\tfrac{1-2\varepsilon}{Q\varepsilon(\varepsilon-1)}\big)\Big],\end{aligned} \tag{19}$$

where in the second line we relax the constraint that $P \in \mathbb{N}$ and expressed more generally $P = \varepsilon Q$, introducing the positive quantity $\varepsilon \le 1$.

Equating the expression in (19) to zero yields the following analytical implicit expression to determine the optimal number of symbols P^* that maximizes the SE of a QSM system with $Q = Tn_T$ spatial-temporal resources and employing an M-ary constellation

$$P^* = \lfloor \varepsilon\, Q \rfloor \ \Big|\ \tfrac{2\varepsilon-1}{(\varepsilon-1)\ln\left(\left(\frac{1-\varepsilon}{\varepsilon}\right)^2 M\right)} = \varepsilon Q \triangleq P, \tag{20}$$

where we emphasize that the quantity on the righthand side of the expression is in fact sought after number of transmitted symbols P.

Recalling that the desired P^* is also the largest possible, (20) implies that

$$P^* = \Big\lfloor \max\Big(\tfrac{2\varepsilon-1}{(\varepsilon-1)\ln(\frac{1-\varepsilon}{\varepsilon})^2 M}\Big)\Big\rfloor, \tag{21}$$

which in turn implies that the optimum ε is such that $\left(\frac{1-\varepsilon}{\varepsilon}\right)^2 M = 1$, *i.e.*, the solution of the quadratic polynomial $(M-1)\varepsilon^2 - 2M\varepsilon + M$, which finally yields, simply

$$P^* = \big\lfloor \tfrac{M-\sqrt{M}}{M-1} Q \big\rfloor = \lfloor \varepsilon^*_M T n_T \rfloor, \tag{22}$$

where we introduced the implicitly-defined optimum gradient $\varepsilon^*_M \triangleq \frac{M-\sqrt{M}}{M-1}$.

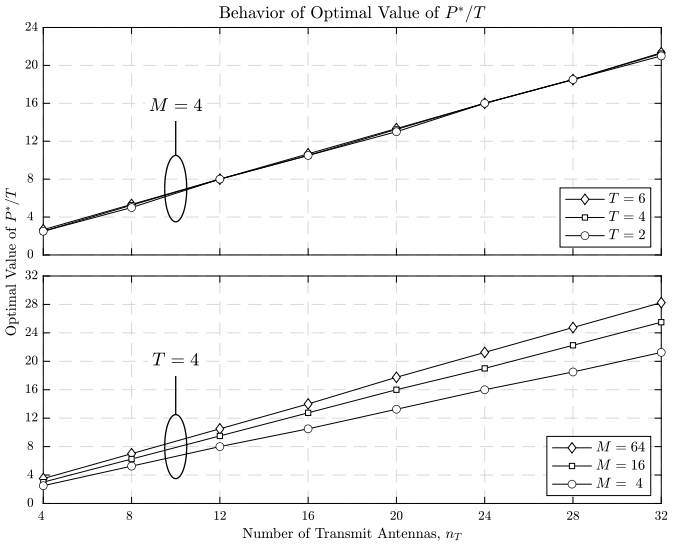


Fig. 3. Effect of T and M on the optimum ratio P^*/T between the number of transmit symbols and epochs.

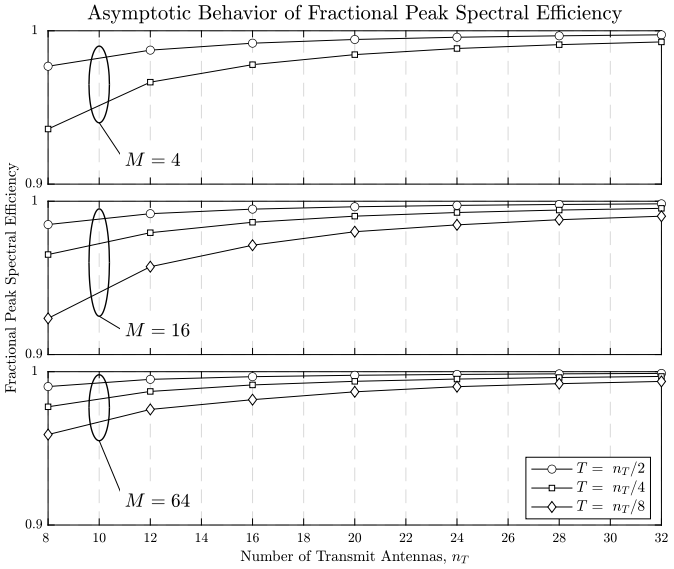


Fig. 4. Behavior of fractional peak spectral efficiency as a function of n_T, for different sizes of T and M.

We emphasize that the elegant result offered in (22) is general for any QSM scheme. From this result it is seen that the optimum ratio P/T that maximizes the SE of the QSM scheme is linear on the number of transmit antennas n_T. In other words, for any given M and n_T, an SE-optimum QSM must be such that P/T scales linearly with n_T, as illustrated and confirmed by the simulation results shown in Fig. 3.

Recall also that QSM dispersion matrices are constructed with basis on STBCs characterized by $T \times T$ square encoding matrices. Consequently, it follows that if P must scale with n_T in order for the QSM to be SE-optimal, so must the size T of the code, in order for the the underlying STBC itself to retain SE-optimality. In other words, (22) also implies that in order to achieve SE-optimality, a QSM scheme conveying M-ary symbols must employ an underlying full-rate STBC of a size that scales proportionally to the number of transmit antennas n_T.

We remark, however, that setting $T = n_T$ is not a scalable proposition, not only because it implies furbishing the transmitter with an equal number of RF chains, which can be prohibitively expensive, but also because it results in fully dense signals, which in turn require also prohibitively complex ML receivers. This observation motivates the comparisons given in Fig. 4, which shows the fraction of the maximum attainable spectral efficiencies ζ^* occurring at P^*, obtained by QSM schemes employing STBCs of different sizes, as a function of n_T and for different M. It can be seen that QSM schemes with T sufficiently large, but still significantly less than n_T, also asymptotically achieve near optimal SE as long as n_T is sufficiently large.

D. Optimal Set Selection of Dispersion Matrix Indices

On top of the coding-gain and SE optimization, and the temporal scalability, there is another mechanism to improve the performance of QSM schemes employing STBCs, namely, to optimize the selection of the index vectors in $\mathcal{K} = \{\boldsymbol{k}_n\}_{n=1}^{N} \in \mathbb{N}^P$ that determine which dispersion matrices are assigned to the real and imaginary parts of each encoded symbol. This is because each index vectors $\boldsymbol{k}_n$ is, according to (10) and (11), associated with different subsets of spatial-temporal resources utilized by the QSM scheme in the transmission of a given set of spatially encoded bits. To illustrate the issue, define the set $\mathcal{K}^*$ of all $\binom{Q}{P}$ distinct index vectors for a given pair (P, Q), and consider the corresponding example compiled in Table I for the case $P = 3$, $T = 2$, and $Q = 8$ ($n_T = 4$, with $Q \triangleq Tn_T$).

Recall also that each dispersion matrix in the transmission of s_p^R or s_p^I uses P given pairs of antennas and time slots, as per (10) and (11), such that for the sake of conciseness we hereafter refer to each pair of one antenna and one time slot simply as a spatial temporal *resource* r_q, defining also for future convenience the set of all available and utilized spatial temporal resources, denoted respectively by $\mathcal{R}^*$ and $\mathcal{R}$. Then, if resources and dispersion matrix indices are represented by a rectangular and a circular nodes, respectively, a bipartite graph such as the one shown in Fig. 5 for the case in question ($i.e., P = 3$, $T = 2$, and $Q = Tn_T = 8$) can be built, in which an edge connecting a circular and a rectangular nodes indicates that the corresponding resource is used by the given dispersion matrix.

As illustrated by the graph, the inclusion of a given index set $\boldsymbol{k}_n$ from $\mathcal{K}^*$ into the set $\mathcal{K}$ is associated with the use of certain resources, occasionally with multiplicity, identified by the graph edges intercepted by the enclosure encircling the corresponding indices. We shall therefore use the notation $\boldsymbol{k}_n \Rightarrow \boldsymbol{r}_n$ to indicate that the index set $\boldsymbol{k}_n$ implies the utilization of the set of resources $\boldsymbol{r}_n$, and $\mu_{\boldsymbol{k}_n}(r_q)$ to denote the multiplicity of the resource r_q in the set $\boldsymbol{k}_n$.

For example, the use of the resources $\boldsymbol{r}_1 = \{2 \times (1,1),\ (1,2),\ (2,1),\ 2 \times (2,2)\}$ results from having $\boldsymbol{k}_1 = [1,\ 2,\ 3]$ in $\mathcal{K}$, such that we may write concisely $\boldsymbol{k}_1 \Rightarrow \boldsymbol{r}_1$, with $\mu_{\boldsymbol{k}_1}(1,1) = \mu_{\boldsymbol{k}_1}(2,2) = 2$. Similarly,

TABLE I
SETS OF ALL POSSIBLE INDEX VECTORS $\mathcal{K}^*$ AND RESOURCES $\mathcal{R}^*$ ($P = 3$, $n_T = 4$, $T = 2$)

Elements of $\mathcal{K}^*$		Elements of $\mathcal{R}^*$	Elements of $\mathcal{K}^*$		Elements of $\mathcal{R}^*$
$\boldsymbol{k}_1$	**[1, 2, 3]**	**(1,1), (2,2), (1,2), (2,1), (1,1), (2,2)**	$\boldsymbol{k}_{29}$	[2, 4, 7]	(1,2), (2,1), (1,2), (2,1), (3,1), (4,2)
$\boldsymbol{k}_2$	**[1, 2, 4]**	**(1,1), (2,2), (1,2), (2,1), (1,2), (2,1)**	$\boldsymbol{k}_{30}$	[2, 4, 8]	(1,2), (2,1), (1,2), (2,1), (3,2), (4,1)
$\boldsymbol{k}_3$	**[1, 2, 5]**	**(1,1), (2,2), (1,2), (2,1), (3,1), (4,2)**	$\boldsymbol{k}_{31}$	[2, 5, 6]	(1,2), (2,1), (3,1), (4,2), (3,2), (4,1)
$\boldsymbol{k}_4$	[1, 2, 6]	(1,1), (2,2), (1,2), (2,1), (3,2), (4,1)	$\boldsymbol{k}_{32}$	[2, 5, 7]	(1,2), (2,1), (3,1), (4,2), (3,1), (4,2)
$\boldsymbol{k}_5$	**[1, 2, 7]**	**(1,1), (2,2), (1,2), (2,1), (3,1), (4,2)**	$\boldsymbol{k}_{33}$	[2, 5, 8]	(1,2), (2,1), (3,1), (4,2), (3,2), (4,1)
$\boldsymbol{k}_6$	**[1, 2, 8]**	**(1,1), (2,2), (1,2), (2,1), (3,2), (4,1)**	$\boldsymbol{k}_{34}$	[2, 6, 7]	(1,2), (2,1), (3,2), (4,1), (3,1), (4,2)
$\boldsymbol{k}_7$	**[1, 3, 4]**	**(1,1), (2,2), (1,1), (2,2), (1,2), (2,1)**	$\boldsymbol{k}_{35}$	**[2, 6, 8]**	**(1,2), (2,1), (3,2), (4,1), (3,2), (4,1)**
$\boldsymbol{k}_8$	**[1, 3, 5]**	**(1,1), (2,2), (1,1), (2,2), (3,1), (4,2)**	$\boldsymbol{k}_{36}$	**[2, 7, 8]**	**(1,2), (2,1), (3,1), (4,2), (3,2), (4,1)**
$\boldsymbol{k}_9$	[1, 3, 6]	(1,1), (2,2), (1,1), (2,2), (3,2), (4,1)	$\boldsymbol{k}_{37}$	**[3, 4, 5]**	**(1,1), (2,2), (1,2), (2,1), (3,1), (4,2)**
$\boldsymbol{k}_{10}$	**[1, 3, 7]**	**(1,1), (2,2), (1,1), (2,2), (3,1), (4,2)**	$\boldsymbol{k}_{38}$	**[3, 4, 6]**	**(1,1), (2,2), (1,2), (2,1), (3,2), (4,1)**
$\boldsymbol{k}_{11}$	**[1, 3, 8]**	**(1,1), (2,2), (1,1), (2,2), (3,2), (4,1)**	$\boldsymbol{k}_{39}$	[3, 4, 7]	(1,1), (2,2), (1,2), (2,1), (3,1), (4,2)
$\boldsymbol{k}_{12}$	[1, 4, 5]	(1,1), (2,2), (1,2), (2,1), (3,1), (4,2)	$\boldsymbol{k}_{40}$	[3, 4, 8]	(1,1), (2,2), (1,2), (2,1), (3,2), (4,1)
$\boldsymbol{k}_{13}$	[1, 4, 6]	(1,1), (2,2), (1,2), (2,1), (3,2), (4,1)	$\boldsymbol{k}_{41}$	**[3, 5, 6]**	**(1,1), (2,2), (3,1), (4,2), (3,2), (4,1)**
$\boldsymbol{k}_{14}$	[1, 4, 7]	(1,1), (2,2), (1,2), (2,1), (3,1), (4,2)	$\boldsymbol{k}_{42}$	**[3, 5, 7]**	**(1,1), (2,2), (3,1), (4,2), (3,1), (4,2)**
$\boldsymbol{k}_{15}$	[1, 4, 8]	(1,1), (2,2), (1,2), (2,1), (3,2), (4,1)	$\boldsymbol{k}_{43}$	[3, 5, 8]	(1,1), (2,2), (3,1), (4,2), (3,2), (4,1)
$\boldsymbol{k}_{16}$	[1, 5, 6]	(1,1), (2,2), (3,1), (4,2), (3,2), (4,1)	$\boldsymbol{k}_{44}$	[3, 6, 7]	(1,1), (2,2), (3,2), (4,1), (3,1), (4,2)
$\boldsymbol{k}_{17}$	[1, 5, 7]	(1,1), (2,2), (3,1), (4,2), (3,1), (4,2)	$\boldsymbol{k}_{45}$	[3, 6, 8]	(1,1), (2,2), (3,2), (4,1), (3,2), (4,1)
$\boldsymbol{k}_{18}$	[1, 5, 8]	(1,1), (2,2), (3,1), (4,2), (3,2), (4,1)	$\boldsymbol{k}_{46}$	[3, 7, 8]	(1,1), (2,2), (3,1), (4,2), (3,2), (4,1)
$\boldsymbol{k}_{19}$	**[1, 6, 7]**	**(1,1), (2,2), (3,2), (4,1), (3,1), (4,2)**	$\boldsymbol{k}_{47}$	**[4, 5, 6]**	**(1,2), (2,1), (3,1), (4,2), (3,2), (4,1)**
$\boldsymbol{k}_{20}$	**[1, 6, 8]**	**(1,1), (2,2), (3,2), (4,1), (3,2), (4,1)**	$\boldsymbol{k}_{48}$	**[4, 5, 7]**	**(1,2), (2,1), (3,1), (4,2), (3,1), (4,2)**
$\boldsymbol{k}_{21}$	**[1, 7, 8]**	**(1,1), (2,2), (3,1), (4,2), (3,2), (4,1)**	$\boldsymbol{k}_{49}$	[4, 5, 8]	(1,2), (2,1), (3,1), (4,2), (3,2), (4,1)
$\boldsymbol{k}_{22}$	**[2, 3, 4]**	**(1,2), (2,1), (1,1), (2,2), (1,2), (2,1)**	$\boldsymbol{k}_{50}$	**[4, 6, 7]**	**(1,2), (2,1), (3,2), (4,1), (3,1), (4,2)**
$\boldsymbol{k}_{23}$	**[2, 3, 5]**	**(1,2), (2,1), (1,1), (2,2), (3,1), (4,2)**	$\boldsymbol{k}_{51}$	**[4, 6, 8]**	**(1,2), (2,1), (3,2), (4,1), (3,2), (4,1)**
$\boldsymbol{k}_{24}$	[2, 3, 6]	(1,2), (2,1), (1,1), (2,2), (3,2), (4,1)	$\boldsymbol{k}_{52}$	**[4, 7, 8]**	**(1,2), (2,1), (3,1), (4,2), (3,2), (4,1)**
$\boldsymbol{k}_{25}$	[2, 3, 7]	(1,2), (2,1), (1,1), (2,2), (3,1), (4,2)	$\boldsymbol{k}_{53}$	**[5, 6, 7]**	**(3,1), (4,2), (3,2), (4,1), (3,1), (4,2)**
$\boldsymbol{k}_{26}$	**[2, 3, 8]**	**(1,2), (2,1), (1,1), (2,2), (3,2), (4,1)**	$\boldsymbol{k}_{54}$	**[5, 6, 8]**	**(3,1), (4,2), (3,2), (4,1), (3,2), (4,1)**
$\boldsymbol{k}_{27}$	**[2, 4, 5]**	**(1,2), (2,1), (1,2), (2,1), (3,1), (4,2)**	$\boldsymbol{k}_{55}$	**[5, 7, 8]**	**(3,1), (4,2), (3,1), (4,2), (3,2), (4,1)**
$\boldsymbol{k}_{28}$	**[2, 4, 6]**	**(1,2), (2,1), (1,2), (2,1), (3,2), (4,1)**	$\boldsymbol{k}_{56}$	**[6, 7, 8]**	**(3,2), (4,1), (3,1), (4,2), (3,2), (4,1)**

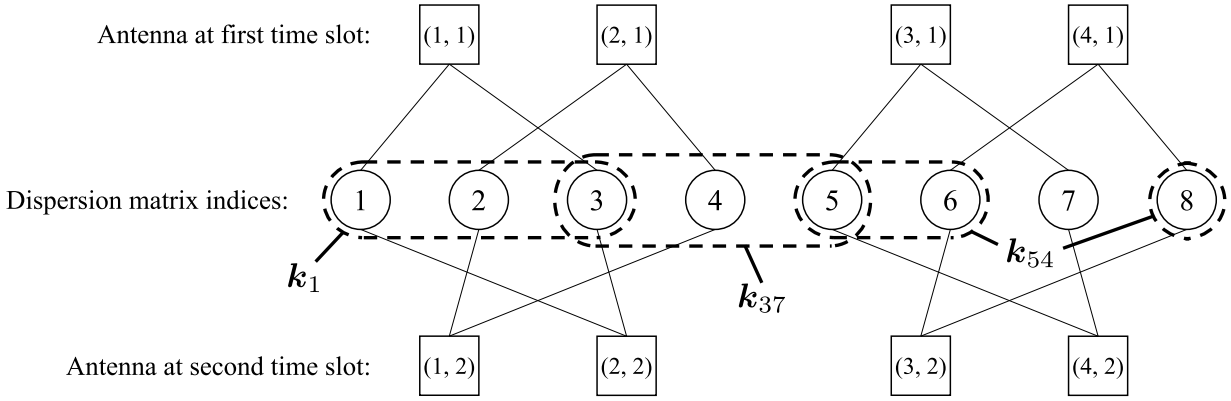


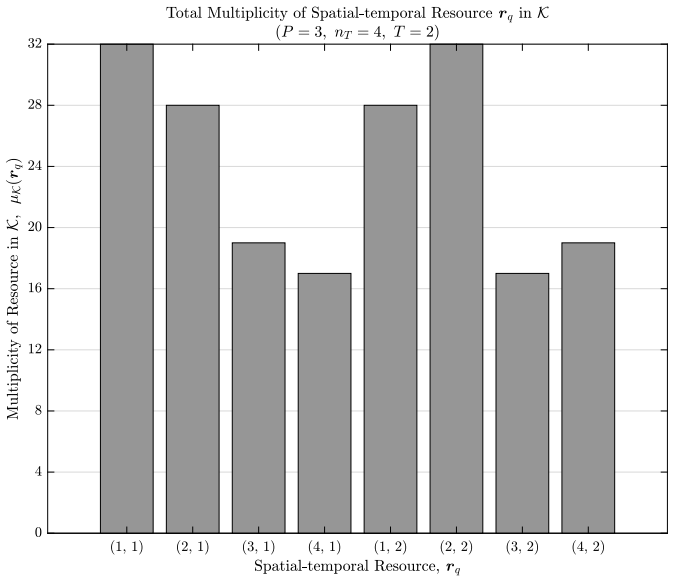
Fig. 5. Bipartite graph representing the spatial-temporal resource usage associated with each index vector $\boldsymbol{k}_n$ for a QSM system with $P = 3$, $T = 2$, and $Q = 8$. The particular examples of $\boldsymbol{k}_1$, $\boldsymbol{k}_{37}$ and $\boldsymbol{k}_{54}$ are explicitly illustrated.

$\boldsymbol{k}_{37} = [3, 4, 5] \Rightarrow \boldsymbol{r}_{37} = \{(1, 1), (1, 2), (2, 1), (2, 2), (3, 1), (4, 2)\}$, and $\boldsymbol{k}_{54} = [5, 6, 8] \Rightarrow \boldsymbol{r}_{54} = \{(3, 1), 2 \times (3, 2), 2 \times (4, 1), (4, 2)\}$, with $\mu_{\boldsymbol{r}_{37}}(3, 2) = \mu_{\boldsymbol{r}_{37}}(4, 1) = 2$.

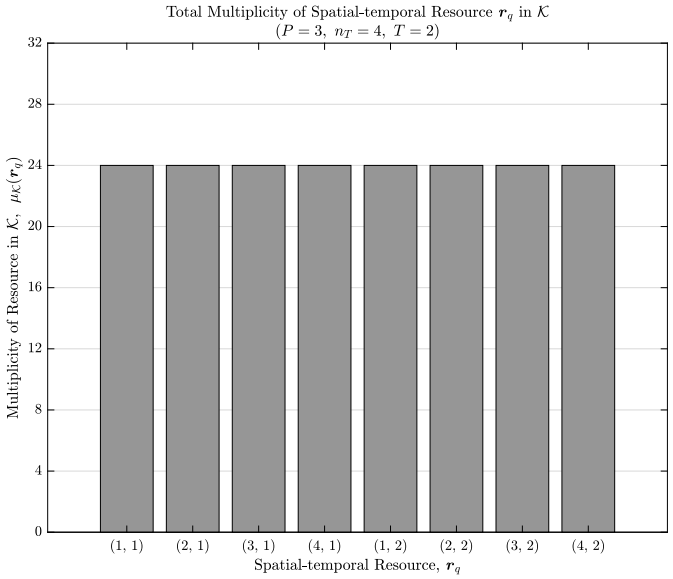
It is evident from all the above that in order to avoid redundancy and uneven utilization of spatial-temporal resources, so as to optimize the performance of QSM schemes [38], the sets of dispersion matrix indices $\mathcal{K}$ (with corresponding resource set $\mathcal{R}$) must satisfy the following conditions: a) no two index vectors $\boldsymbol{k}_n$ and $\boldsymbol{k}_m$ in the set can be equal ($i.e., \boldsymbol{k}_n \neq \boldsymbol{k}_m, \forall\, n \neq m$); b) no two elements in each index vector can be equal $\big(i.e., [\boldsymbol{k}_n]_i \neq [\boldsymbol{k}_n]_j, \forall\, \boldsymbol{k}_n \text{ and } i \neq j\big)$; c) the utilization of all resources available must be ensured $\big(i.e., \mu_{\mathcal{K}}(r_q) > 0 \;\forall\; r_q \in \mathcal{R}\big)$; d) all resources are utilized as often $\big(i.e., \mu_{\mathcal{K}}(r_1) = \cdots = \mu_{\mathcal{K}}(r_Q)\big)$, and finally e) the cardinality of the set must be a power of 2 in order to enable the encoding of codewords $\big(i.e., N = |\mathcal{K}| = \lfloor \binom{Q}{P} \rfloor_{2^\times}\big)$.

As an example, Table I highlights the set of index vectors $\mathcal{K} = \{\boldsymbol{k}_1, \cdots, \boldsymbol{k}_3,\ \boldsymbol{k}_5, \cdots, \boldsymbol{k}_8,\ \boldsymbol{k}_{10}, \boldsymbol{k}_{11},\ \boldsymbol{k}_{19}, \cdots, \boldsymbol{k}_{23},\ \boldsymbol{k}_{26}, \cdots, \boldsymbol{k}_{28},\ \boldsymbol{k}_{35}, \cdots, \boldsymbol{k}_{38},\ \boldsymbol{k}_{41},\ \boldsymbol{k}_{42},\ \boldsymbol{k}_{47},\ \boldsymbol{k}_{48},\ \boldsymbol{k}_{50}, \cdots, \boldsymbol{k}_{56}\}$. The reader can verify that by this choice of $\mathcal{K}$, all resources in the associated set $\mathcal{R}$ have multiplicity 24. In contrast, a naive truncation of the first 32 index vectors in Table I, $i.e.$, $\mathcal{K} = \{\boldsymbol{k}_1, \cdots, \boldsymbol{k}_{32}\}$ as suggested $e.g.$ in [25], leads to an uneven utilization pattern in which $\mu_{\mathcal{K}}(1, 1) = \mu_{\mathcal{K}}(2, 2) = 32$, $\mu_{\mathcal{K}}(1, 2) = \mu_{\mathcal{K}}(2, 1) = 28$, $\mu_{\mathcal{K}}(3, 1) = \mu_{\mathcal{K}}(4, 2) = 19$ and $\mu_{\mathcal{K}}(3, 2) = \mu_{\mathcal{K}}(4, 1) = 17$, which is obviously sub-optimum as it leads to antennas 1 and 2 being used far more often than antennas 3 and 4.

To see this effect clearly, we provide a visual representation in Fig. 6a where the histogram exhibits the allocation pattern of the spatial-temporal resources of the naively selected set, as per the SotA method.



(a) SotA naive allocation.



(b) Proposed optimal allocation.

Fig. 6. Comparison of the spatial-temporal resource allocation patterns for the naive and proposed schemes.

The problem of selecting the optimum set $\mathcal{K}$ described above and illustrated in Fig. 6b, relates to a classic problem in combinatorics graph theory [39] known as the Vertex Cover Problem. In the context hereby, however, the problem has the additional difficulties that: a) the graph in question is bipartite, b) coverage with equal multiplicity is required rather than just full coverage, and c) nodes must be selected in subsets of P at a time. Due to these peculiarities, the problem itself is, to the best of our knowledge, original and cannot be solved by known variations of the Vertex Cover algorithm. Fortunately, the highly symmetric structure of the associated bipartite graph can be exploited to design an efficient algorithm to solve the selection problem at hand.

Algorithm 1 Greedy Construction of Optimal Set of Index Vectors $\mathcal{K}$

Internal Parameters: Number of resources $Q = T \cdot n_T$ and set of all possible indices $\mathcal{K}^*$.

Inputs: Number of symbols P, of transmit antennas n_T and dimension T of FDFR STBC.

Outputs: Optimized set of index vectors $\mathcal{K}$.

Choose a random seed $n \in \left\{1, \cdots, \binom{Q}{P}\right\}$ and start with $\mathcal{K} = \emptyset$;
while $|\mathcal{K}| \neq \lfloor \binom{Q}{P} \rfloor_{2\times}$ **do**
  Insert $\boldsymbol{k}_n$ into the set $\mathcal{K}$ of selected index vectors;
  Sort all indices $k \in \{1, \cdots, Q\}$ in ascending order of their multiplicities in $\mathcal{K}$;
  Set $D = P$ and construct/clear the empty set $\tilde{\mathcal{K}} = \emptyset$ of candidate index vectors;
  while $|\tilde{\mathcal{K}}| = 0$ **do**
    Construct a list $\boldsymbol{\kappa}$ of candidate indices with the D lowest multiplicities in $\mathcal{K}$;
    Construct the set $\tilde{\mathcal{K}}$ of all $\binom{D}{P}$ index vectors $\tilde{\boldsymbol{k}}_m$ with indices in $\boldsymbol{\kappa}$;
    Remove from $\tilde{\mathcal{K}}$ all index vectors already in $\mathcal{K}$;
    if $|\tilde{\mathcal{K}}| = 0$ **then**
      Increment D by 1;
    end if
  end while
  Select next $n \in \left\{1, \cdots, \binom{Q}{P}\right\}$ as the position of the first index vector $\tilde{\boldsymbol{k}}_m$ of $\tilde{\mathcal{K}}$ in $\mathcal{K}^*$;
end while

To that end, let us commit a slight abuse of notation and define the multiplicity of a dispersion matrix index[5] q in the set $\mathcal{K}$ as $\mu_{\mathcal{K}}(q)$. Then, by virtue of the symmetry of the graph (see Fig. 5), a solution $\mathcal{K}$ in which $\mu_{\mathcal{K}}(1) = \cdots = \mu_{\mathcal{K}}(Q)$ implies a solution $\mathcal{R}$ in which each of the spatial temporal resources $\{(1,1),(1,2),(2,1),(2,2),(3,1),(4,2)\}$ have the same multiplicity. Consequently, the problem can be solved efficiently by the greedy selection of indices, as described in Algorithm 1.

Notice that while the examples given above are for the minimal case of $T = 2$ as in the G-QSM or the EDA-QSM, it is obvious that the algorithm is also valid for any general T.

This holds because the key properties of the underlying bipartite graph enabling the proposed algorithm, *i.e.,* symmetry and disjointedness, are still retained for dispersion matrices for $T > 2$ constructed following (14) and (15).[6] Following this, the corresponding bipartite graph illustrated in Fig. 5 is

[5]Notice that there is no ambiguity with the definition of multiplicity of resources because dispersion matrix indices are single numbers, while spatial temporal resources are pairs.

[6]To see this, inspect (14), where each auxiliary matrix $\mathbf{C}_i$ and $\mathbf{D}_i$ is a $T \times T$ sparse matrix obtained from a cyclic rotation of a diagonal matrix containing only T non-zero elements of $\mathbf{R}$, and (15) whose kronecker products produces the disjointedness in the associated dispersion matrices and the graph.

merely expanded to a similar graph with $T\cdot n_T$ index (circular) nodes and $T\cdot n_T$ resource (rectangular) nodes, with each q resource node connected to T index nodes and vice versa. As a result, the greedy strategy described earlier remains valid, as evidenced by the fact that Algorithm 1 applies to general T.

E. Optimized Scalable Quadrature Spatial Modulation

Finally, the contributions of the previous Subsections III-B to III-D are combined, and proposed as the optimized scalable QSM (OS-QSM) transmitter design, which is optimized under SE, coding-gain, and resource allocation, while scalable in both the spatial and temporal domain. For the convenience of the reader, we summarize the structure of the proposed OS-QSM transmitter design in Algorithm 2.

Algorithm 2 Proposed OS-QSM Signal Generation

Internal Parameters: Number of symbols P, transmit antennas n_T, time slots T and spatial-temporal resources $Q = T\cdot n_T$; and cardinalities $M = |\mathcal{S}|$, and $N = \lfloor \binom{Q}{P} \rfloor_{2^\times}$.

Global Quantities: Symbol constellation $\mathcal{S}$, optimum lattice generating matrix $\mathbf{R}$, sets of dispersion matrices $\mathcal{A} = \{\mathbf{A}_q\}_{q=1}^{Q}$ and $\mathcal{B} = \{\mathbf{B}_q\}_{q=1}^{Q}$ with $\mathbf{A}_q$ and $\mathbf{B}_q$ as in (15), set of index vectors $\mathcal{K}$ obtained from Algorithm 1.

Input: Information bit sequence $\boldsymbol{b} = [\boldsymbol{b}^R, \boldsymbol{b}^I, \boldsymbol{b}^S]$;
Outputs: Transmitted signal $\mathbf{X}$.

1: Select vector $\boldsymbol{k}^R$ as the $([\boldsymbol{b}^R]_{(10)}+1)$-th vector in $\mathcal{K}$;
2: Select vector $\boldsymbol{k}^I$ as the $([\boldsymbol{b}^I]_{(10)}+1)$-th vector in $\mathcal{K}$;
3: Assign the bits $\boldsymbol{b}^S$ to P symbols $\{s_1, \cdots, s_P\}$ selected from $\mathcal{S}$, with $s_p = s_p^R + j s_p^I$;
4: Construct $\mathbf{X} = \sum_{p=1}^{P} (s_p^R \mathbf{A}_{k_p^R} + s_p^I \mathbf{B}_{k_p^I})$ as per (2);

It should be mentioned that the aforementioned contributions may be integrated, jointly or separately, into other general SM/QSM schemes and is not limited to our specific transmitter design, subject to an appropriate modification. However such modifications and applications are out of scope of this paper.

IV. Proposed Receiver Design

A. Sparse Formulation of QSM Receivers

Together, Algorithms 1 and 2 introduced above demonstrate that the design of OS-QSM transmitters is possible and tractable. There is, however, no true scalability without feasibility, such that in order to complete the task it is also necessary to show that the proposed OS-QSM design is effectively decodable at reasonable complexity.

To put the challenge into context, for given P, M, T and n_T, with $Q = T\cdot n_T$, an ML receiver would have to go through $(\lfloor \binom{Q}{P} \rfloor_{2^\times})^2 \cdot M^P$ combinations of symbols and selected spatial-temporal resources in order to detect a sequence of $2\cdot \lfloor \log_2 \binom{Q}{P} \rfloor + P\cdot \log_2 M$ bits. That means that even for the minimal setting of $T=3$, $P=3$ and $M=4$, a system with $n_T = 6$ transmit antennas would require the receiver to go through $(\lfloor \binom{18}{3} \rfloor_{2^\times})^2 \cdot (4)^2 =$16,777,216 combinations in order to decode the corresponding $\log_2$(16,777,216) $= 24$ bits. In other words, ML decoding is highly impractical in QSM systems, especially in the context of massive MIMO systems.

We emphasize that this challenge applies not only to the OS-QSM scheme of Subsection III, but also to current SotA QSM methods such as those in [23]–[25], as the example given above is for $T=2$, which is the size of the core codes used in the latter. We furthermore stress that the utilization of SD receivers is also not viable in scaled cases, because the nature of tree search algorithms still requires excessive computational complexity in large systems, as shown in [29]–[31]. Finally, we also remark that since convenient properties such as fast-decodability and block-diagonality are known not to be retainable without sacrifice of optimality for STBC of arbitrary size [33], [34], a scalable detector for QSM schemes cannot rely on such features.

In light of the above, we introduce hereafter a new detection method for QSM schemes which relies neither on tree-search, nor on specific properties of STBCs, and which is completely independent of the infeasible combinatorial factor $\lfloor \binom{Q}{P} \rfloor_{2^\times}$. In addition, given prior information on the encoding construction, the proposed decoder is valid to detect any QSM signal.

The core idea of our approach is to take full advantage of a sparse representation of QSM signals over the entire channel ($i.e.$, for all spatial temporal resources available), assumed known at the receiver. The proposed decoding method then leverages the iterative shrinkage thresholding algorithm (ISTA) to greedily extract symbol and dispersion index estimates, resulting in significantly lower complexities compared to ML and SD-based methods. To that end, first combine (1) and (2), and consider the vectorized form of the QSM received signal

$$\begin{aligned}\mathbf{y} \triangleq \mathrm{vec}(\mathbf{Y}) &= \overbrace{(\mathbf{I}_T \otimes \mathbf{H})}^{\triangleq \boldsymbol{\Phi}_H}(\boldsymbol{\Xi}_A \mathbf{u}^R + \boldsymbol{\Xi}_B \mathbf{u}^I) + \overbrace{\mathrm{vec}(\mathbf{V})}^{\triangleq \mathbf{v}} \\ &= \boldsymbol{\Phi}_H \cdot (\boldsymbol{\Xi}_A \mathbf{u}^R + \boldsymbol{\Xi}_B \mathbf{u}^I) + \mathbf{v} \ \in \mathbb{C}^{Tn_R \times 1},\end{aligned} \tag{23}$$

where we implicitly defined the block-diagonal channel matrix $\boldsymbol{\Phi}_H$ and vectorized noise $\mathbf{v}$; the dispersion matrices in $\mathcal{A}$ and $\mathcal{B}$ are also vectorized into $\mathbf{a}_q \triangleq \mathrm{vec}(\mathbf{A}_q)$ and $\mathbf{b}_q \triangleq \mathrm{vec}(\mathbf{B}_q)$ and concatenated into $\boldsymbol{\Xi}_A \triangleq [\mathbf{a}_1, \cdots, \mathbf{a}_Q] \in \mathbb{C}^{Q\times Q}$ and $\boldsymbol{\Xi}_B = [\mathbf{b}_1, \cdots, \mathbf{b}_Q] \in \mathbb{C}^{Q\times Q}$, respectively; and $\mathbf{u}^R \in \mathbb{R}^{Q\times 1}$ (respectively $\mathbf{u}^I \in \mathbb{R}^{Q\times 1}$) is set to zero everywhere, except for its elements of indices $\boldsymbol{k}^R \in \mathcal{K}$ (respectively $\boldsymbol{k}^I \in \mathcal{K}$), which are set to $\{s_1^R, \cdots, s_P^R\}$ (respectively $\{s_1^I, \cdots, s_P^I\}$).

Equation (23) can be further simplified by defining the combined and real-imaginary decoupled information and noise vectors

$$\boldsymbol{u} \triangleq [\,\mathrm{u}_1^R, \mathrm{u}_1^I, \cdots, \mathrm{u}_Q^R, \mathrm{u}_Q^I\,]^\mathsf{T} \in \mathbb{R}^{2Q\times 1}, \tag{24}$$

$$\boldsymbol{v} \triangleq [v_1^R, v_1^I, \cdots, v_{Tn_R}^R, v_{Tn_R}^I]^\mathsf{T} \in \mathbb{R}^{2Q\times 1}, \tag{25}$$

as well as the decoupled versions of $\mathbf{a}_q$ and $\mathbf{b}_q$, namely

$$\boldsymbol{a}_q \triangleq [\,\mathrm{a}_{q_1}^R, \mathrm{a}_{q_1}^I, \cdots, \mathrm{a}_{qQ}^R, \mathrm{a}_{qQ}^I\,]^\mathsf{T} \in \mathbb{R}^{2Q\times 1}, \tag{26a}$$

$$\boldsymbol{b}_q \triangleq [\,\mathrm{b}_{q_1}^R, \mathrm{b}_{q_1}^I, \cdots, \mathrm{b}_{qQ}^R, \mathrm{b}_{qQ}^I\,]^\mathsf{T} \in \mathbb{R}^{2Q\times 1}, \tag{26b}$$

which in turn can be combined into a single dispersion matrix $\boldsymbol{\Psi}_D \in \mathbb{R}^{2Q\times 2Q}$, namely

$$\boldsymbol{\Psi}_D \triangleq [\,\boldsymbol{a}_1, \boldsymbol{b}_1, \boldsymbol{a}_2, \boldsymbol{b}_2, \cdots, \boldsymbol{a}_Q, \boldsymbol{b}_Q\,], \tag{27}$$

such that the vectorized model of (23) can be rewritten as

$$\begin{aligned}\boldsymbol{y} &= \check{\boldsymbol{\Phi}}_H \boldsymbol{\Psi}_D \cdot \boldsymbol{u} + \boldsymbol{v} = \boldsymbol{\Phi}_H \cdot \boldsymbol{\Psi}_D \cdot \boldsymbol{u} + \boldsymbol{v} \\ &= \boldsymbol{G} \cdot \boldsymbol{u} + \boldsymbol{v} \ \in \mathbb{R}^{2Tn_R \times 1}, \end{aligned} \tag{28}$$

where $\check{\boldsymbol{\Phi}}_H$ is the quadrature-operated block diagonal channel matrix $\boldsymbol{\Phi}_H$, which we implicitly relabeled $\boldsymbol{\Phi}_H$, as with the effective channel matrix $\boldsymbol{G} \triangleq \boldsymbol{\Phi}_H \cdot \boldsymbol{\Psi}_D$, for future convenience.

To elaborate on (28) with an example, consider a system with $P = 3$, $T = 2$ and $n_T = 4$, and assume that for a particular bit sequence $\boldsymbol{b} = [\boldsymbol{b}^R, \boldsymbol{b}^I, \boldsymbol{b}^S]$, the selected index vectors are given by $\boldsymbol{k}^R = \boldsymbol{k}_{10} = [1,\ 3,\ 7]$ and $\boldsymbol{k}^I = \boldsymbol{k}_{47} = [4,\ 5,\ 7]$. Then, the corresponding combined information vector becomes $\boldsymbol{u} = [s_1^R,\ 0,\ 0,\ 0,\ s_2^R,\ 0,\ 0,\ s_1^I,\ 0,\ s_2^I,\ 0,\ 0,\ s_3^R,\ s_3^I,\ 0,\ 0]^{\mathsf{T}}$. Notice that while $\boldsymbol{u}$ carries in the entries s_p^R and s_P^I the $P \log_2 M$ bits corresponding to the $\boldsymbol{b}^S$ subsequence, the remaining $2 \log_2 N$ bits corresponding to the subsequences $\boldsymbol{b}^R$ and $\boldsymbol{b}^I$ are encoded merely by positions of non-zero elements in $\boldsymbol{u}$, regardless of what the values of s_p^R and s_P^I might be, which suggests that the detection of $\boldsymbol{b}^S$ could be done separately from that of $\boldsymbol{b}^R$ and $\boldsymbol{b}^I$.

In principle, the latter feature could be utilized to design an SD receiver for the OS-QSM method proposed above, similarly to how block-separability was exploited in [25] to do so for the EDA-QSM scheme. The problem with that idea is, of course, the prohibitively large number of combinations how the $2P$ elements of the decoupled symbol vector $\boldsymbol{s} = [s_1^R, s_1^I, \cdots, s_P^R, s_P^I]$ can be placed among the $2Q$ entries of $\boldsymbol{u}$. In order to circumvent this challenge, we instead seek to exploit the facts demonstrated in Subsection III-C, namely, that: a) the optimum number P^* of symbols maximizing SE is a fraction of the total spatial-temporal resources $Q = T \cdot n_T$, as per (20); and b) that in large-scale systems with $n_T \gg 1$, a significantly smaller block size T suffices to asymptotically achieve SE optimality, as shown in Fig. 4. Together, these facts imply that the sparsity of $\boldsymbol{u}$ becomes increasingly more prominent in large-scale SE-optimal QSM schemes, which favors sparse recovery algorithms [35], [36]. It is also evident from the inspection of (28) that the matrices $\boldsymbol{\Phi}_H$ and $\boldsymbol{\Psi}_D$ can be respectively interpreted as the sensing and dictionary matrices typical of compressive sensing (CS) models [40], such that recent progress on sparse [41], [42] and discrete-aware receivers [43] can be leveraged.

Taking into account the focus on scalability, which at the receiver side translates to controlling complexity, two suitable candidate methods to be applied for OS-QSM demodulation are the generalized approximate message-passing (GAMP) algorithm [44]–[46], and the iterative shrinkage thresholding algorithm (ISTA) [35], [36], both of which possess quadratic complexity on the size $2Tn_T$ of the signal vector $\boldsymbol{u}$. It is well-known, however, that the GAMP algorithm relies on the particular structure of measurement matrix and the independence of the received signal, which in the case of QSM cannot be generally assumed, as a direct consequence of the utilization of STBCs in the dispersion matrices. In the absence of the required conditions, GAMP receivers yield poor performance, characterized by error-floors at high SNRs[7] [47], [48].

Motivated by this fact, we therefore choose to follow a ISTA-based approach in the design of a low-complexity demodulator for QSM systems, which is described in the sequel. In particular, we introduce an algorithm to detect QSM signals, which is based on a purpose-built variation of ISTA that incorporates modifications both on the thresholding function and on the index vector estimation process, specifically to QSM detection.

B. Greedy Boxed ISTA-Based QSM Decoder

Consider the standard ISTA recursion [35],

$$\hat{\boldsymbol{u}}^{(\eta+1)} = \Lambda\Big(\hat{\boldsymbol{u}}^{(\eta)} + \frac{1}{\alpha}\boldsymbol{G}^{\mathsf{T}}\big(\boldsymbol{y} - \boldsymbol{G}\,\hat{\boldsymbol{u}}^{(\eta)}\big); \frac{\lambda}{2\alpha}\Big), \tag{29}$$

where $\hat{\boldsymbol{u}}^{(\eta)}$ is the estimate of $\boldsymbol{u}$ at the η-th iteration, $\alpha = \text{maxeig}(\boldsymbol{G}^{\mathsf{T}}\boldsymbol{G})$ is the shrinkage step-size,[8] λ is the threshold factor, and $\Lambda(s;\tau)$ is the soft-thresholding function.

A first meaningful modification to such standard ISTA recursions in (29) is to account for the fact that the symbols in the real-valued projected constellation $\mathcal{S}_R \triangleq \mathcal{P}_{\mathbb{R}}(\mathcal{S})$ are finite, such that in addition to the lower limit τ in the vicinity of the origin used to enforce sparsity in the solution, an upper limit $\max(\mathcal{S}_R)$ can be introduced into the thresholding function.

In other words, for the case at hand we replace ISTA's standard soft-thresholding function $\Lambda(s;\tau)$ by a hard-thresholding function leading to the "boxed" hard-thresholding function $\Pi(s;\tau)$ illustrated in Fig. 7 and defined by

$$\Pi(s;\tau) \triangleq \begin{cases} \min(\mathcal{S}_R) & s \le \min(\mathcal{S}_R), \\ s & \min(\mathcal{S}_R) \le s \le -\tau, \\ 0 & |s| \le \tau, \\ s & \tau \le s \le \max(\mathcal{S}_R), \\ \max(\mathcal{S}_R) & \max(\mathcal{S}_R) \le s. \end{cases} \tag{30}$$

Incorporating this modification yields the boxed-hard ISTA (BH-ISTA) receiver described by

$$\hat{\boldsymbol{u}}^{(\eta+1)} = \Pi\Big(\hat{\boldsymbol{u}}^{(\eta)} + \frac{1}{\alpha}\boldsymbol{G}^{\mathsf{T}}\big(\boldsymbol{y} - \boldsymbol{G}\,\hat{\boldsymbol{u}}^{(\eta)}\big); \frac{\lambda}{2\alpha}\Big). \tag{31}$$

Notice that the computational cost of repeatedly evaluating (31) is dominated by the term $\boldsymbol{G}\,\hat{\boldsymbol{u}}^{(\eta)}$, therefore quadratic on the number of non-zero entries ($i.e.$, ℓ_0-norm) of $\hat{\boldsymbol{u}}^{(\eta)}$, which reduces with the iterations η, as illustrated in Fig. 8. In particular, Fig. 8a shows a comparison of the convergence of $|\hat{\boldsymbol{u}}^{(\eta)}|_0$ as a function of η for various values of threshold parameter τ, with $\hat{\boldsymbol{u}}^{(\eta)}$ obtained both from (29) and (31), $i.e.$, via conventional and BH-ISTA, respectively, which for convenience will be hereafter denoted $\hat{\boldsymbol{u}}_{\Pi}^{(\eta)}$ and $\hat{\boldsymbol{u}}_{\Lambda}^{(\eta)}$.

It can in fact be seen that as a result of boxing and hard-thresholding, $|\hat{\boldsymbol{u}}_{\Pi}^{(0)}|_0 < |\hat{\boldsymbol{u}}_{\Lambda}^{(0)}|_0$, such that the expected order of complexity associated with evaluating (29) is lower than

[7] A GAMP receiver for QSM systems was also developed, implemented and used to confirm its shortcoming.

[8] The actual requirement is that $\alpha > \text{maxeig}(\boldsymbol{G}^{\mathsf{T}}\boldsymbol{G})$, however, we will assume the minimum step-size, which is sufficient.

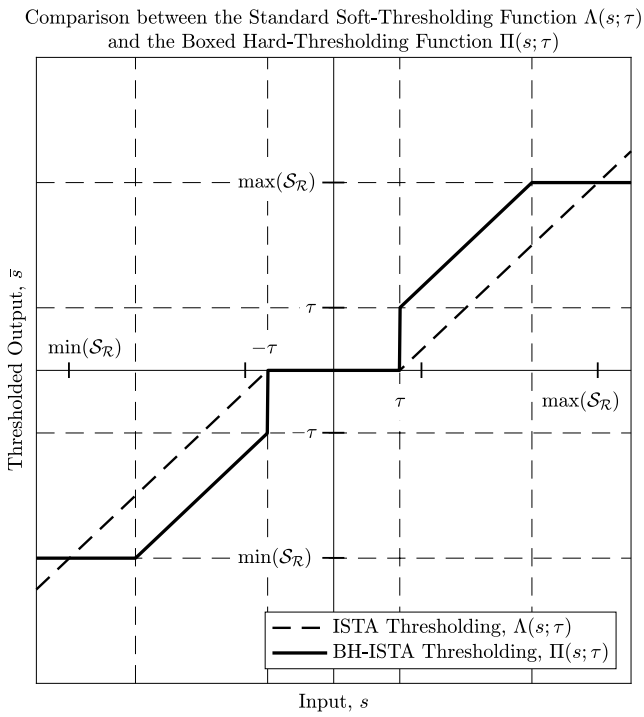
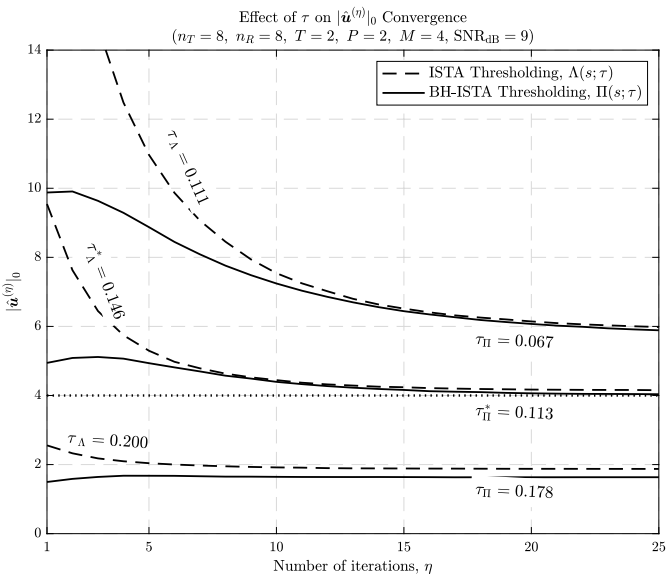
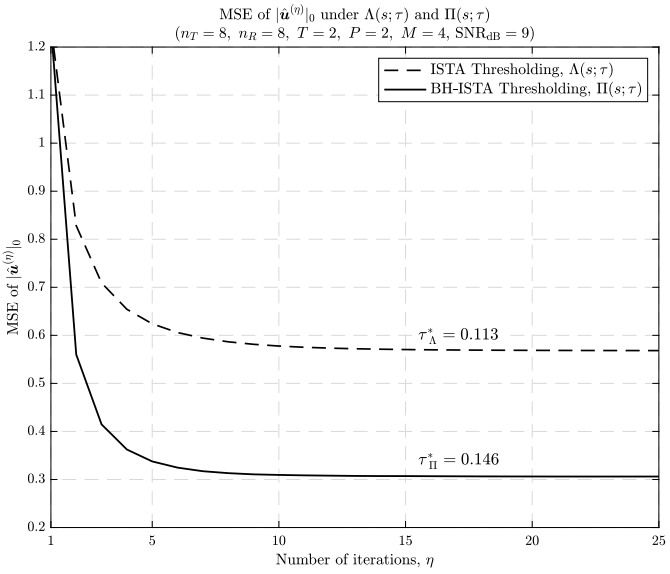


Fig. 7. Comparison of ISTA thresholding and BH-ISTA thresholding functions $\Lambda(s;\tau)$ as per [35], and $\Pi(s;\tau)$, as per (30).



(a) Sparsity convergence with various threshold values.



(b) MSE convergence with optimal threshold values.

Fig. 8. Convergence of $\hat{\boldsymbol{u}}^{(\eta)}_\Pi$ and $\hat{\boldsymbol{u}}^{(\eta)}_\Lambda$, as per (29) and (31), respectively, as a function of iterations η.

that of evaluating (31), which can be bounded both below and above by the lower- and upper-limits $\big(\mathcal{O}(4P^2)$ and $\mathcal{O}(4Q^2)\big)$. More details will be given in Section IV-C. In turn, Fig. 8b shows that the mean-squared error (MSE) obtained with the proposed BH-ISTA approach is better than that obtained with conventional ISTA, which illustrates the effectiveness of the boxed and hard-thresholding modification here proposed for the demodulation of QSM signals.

It is left for us to address, however, how the bits associated with the choices of dispersion matrix indices $\{\boldsymbol{k}^R, \boldsymbol{k}^I\} \in \mathcal{K}$ can be efficiently detected. To that end, another addition is introduced to the ISTA-based detector, namely, a greedy hard-detection procedure for each symbol recovered, with a concomitant update of (31), which can be described as follows.

Let us consider that multiple runs of the BH-ISTA iterations described by (31) are performed, such that prior to the m-th run a modification is made to $\boldsymbol{y}$, $\boldsymbol{G}$ and $\boldsymbol{u}$, which can be expressed by rewriting (31) as

$$\hat{\boldsymbol{u}}_m^{(\eta+1)} = \Pi\Big(\hat{\boldsymbol{u}}_m^{(\eta)} + \frac{1}{\alpha}\boldsymbol{G}_m^{\mathsf{T}}\big(\boldsymbol{y}_m - \boldsymbol{G}_m\,\hat{\boldsymbol{u}}_m^{(\eta)}\big); \frac{\lambda}{2\alpha}\Big), \quad (32)$$

where we convene that for the fist run ($m = 1$) we set $\boldsymbol{y}_1 = \boldsymbol{y}$, $\boldsymbol{G}_1 = \boldsymbol{G}$ and $\hat{\boldsymbol{u}}_1^{(1)} = \boldsymbol{0}_{2Q}$.

Let η^* be the last iteration of the m-th run of the latter estimator, with its corresponding outcome denoted by $\hat{\boldsymbol{u}}_m^{(\eta^*)}$. And finally, let $\tilde{s}_{\hat{q}_m}$ be the entry of $\hat{\boldsymbol{u}}_m^{(\eta^*)}$ with the largest amplitude, whose position is denoted by $\hat{q}_m$, such that we may write

$$\tilde{s}_{\hat{q}_m} = \Big\{[\hat{\boldsymbol{u}}_m^{(\eta^*)}]_{\hat{q}_m} \Big| \, |[\hat{\boldsymbol{u}}_m^{(\eta^*)}]_{\hat{q}_m}| > |[\hat{\boldsymbol{u}}_m^{(\eta^*)}]_\ell|, \forall \ell \in \{1, \cdots, 2Q\}\Big\}, \quad (33)$$

where $[\boldsymbol{x}]_\ell$ denotes the ℓ-th element of a generic vector $\boldsymbol{x}$.

We emphasize that in the greedy procedure summarized by (33), two distinct pieces of information on the bit sequence $\boldsymbol{b}$ are obtained, namely, a soft estimate $\tilde{s}_{\hat{q}_m}$ of one of the modulated symbols $\{s_p^R, s_p^I\} \in \mathcal{S}_R$, and a hard estimate $\hat{q}_m$ of one of the indices contained in the selected index sets $\{\boldsymbol{k}^R, \boldsymbol{k}^I\} \in \mathcal{K}$. In possession of such information, the following steps are then executed in order to produce the modified quantities required to perform the next run of the BH-ISTA recursion described by (32).

First, a hard-detected version of $\tilde{s}_m$ is obtained by projecting in onto $\mathcal{S}_R$, that is

$$\hat{s}_{\hat{q}_m} = \mathcal{P}_{\mathcal{S}_R}(\tilde{s}_{\hat{q}_m}). \quad (34)$$

Then, the remaining quantities are updated following

$$\hat{\boldsymbol{u}}_{m+1}^{(1)} = \big(\boldsymbol{I}_{2Q} - \mathrm{diag}(\boldsymbol{e}_{\hat{q}_m})\big)\hat{\boldsymbol{u}}_m^{(\eta^*)}, \tag{35a}$$

$$\boldsymbol{y}_{m+1} = \boldsymbol{y}_m - \boldsymbol{G}_m \cdot \boldsymbol{e}_{\hat{q}_m}\hat{s}_{\hat{q}_m}, \tag{35b}$$

$$\boldsymbol{G}_{m+1} = \boldsymbol{G}_m\big(\boldsymbol{I}_{2Q} - \mathrm{diag}(\boldsymbol{e}_{\hat{q}_m})\big), \tag{35c}$$

where $\boldsymbol{I}_{2Q}$ is an identity matrix and $\boldsymbol{e}_{\hat{q}_m}$ its $\hat{q}_m$-th column.

Recall that due to the quadrature-decomposed structure of the sparse vector $\boldsymbol{u}$, all odd index estimates $\hat{q}_m$ correspond to the real parts of modulated symbols, while even $\hat{q}_m$ correspond to the imaginary parts, respectively. It is therefore sensible that, as (32) through (35) are evaluated iteratively, the obtained index estimates $\{\hat{q}_1, \hat{q}_2, \cdots, \hat{q}_m\}$ be split and collected accordingly into the subsequences

$$\hat{\boldsymbol{q}}_m^R \triangleq \big\{\hat{q}_m \,\big|\, \mathrm{mod}(\hat{q}_m, 2) = 1, \forall\, m\big\}, \tag{36a}$$

$$\hat{\boldsymbol{q}}_m^I \triangleq \big\{\hat{q}_m \,\big|\, \mathrm{mod}(\hat{q}_m, 2) = 0, \forall\, m\big\}, \tag{36b}$$

where $\mathrm{mod}(x, 2)$ denotes the modulo-2 operation onto x.

If there are no errors during the detection process, after exactly $m = 2P$ runs, the sequences $\{\hat{\boldsymbol{q}}_m^R, \hat{\boldsymbol{q}}_m^I\}$ can be perfectly mapped to $\{\boldsymbol{k}^R, \boldsymbol{k}^I\}$, in particular via

$$\hat{\boldsymbol{k}}^R = \frac{1}{2}(\hat{\boldsymbol{q}}_m^R + 1) \quad \text{and} \quad \hat{\boldsymbol{k}}^I = \frac{1}{2}\hat{\boldsymbol{q}}_m^I, \tag{37}$$

such that procedure comes to a stop.

More generally, however, errors may occur, such that either $\hat{\boldsymbol{q}}_m^R$ or $\hat{\boldsymbol{q}}_m^I$, or both, contain incorrect indices even with cardinality P. In such cases, the procedure continues until both subsequences contains the first P-tuple of indices included in the dispersion matrix index vector set $\mathcal{K}$, at which point a modification of the update equations is required, which can be described as follows.

Let $\mathcal{P}_{\mathcal{K}}(\boldsymbol{q})$ denote the projection of a sequence $\boldsymbol{q}$ onto the set $\mathcal{K}$, such that either a sequence $\boldsymbol{k} \in \mathcal{K}$ or the empty set $\emptyset$ is returned by the projection, depending on whether or not $\boldsymbol{q}$ contains within it a sequence from $\mathcal{K}$. If multiple valid $\boldsymbol{k} \in \mathcal{K}$ exist in the combination of elements in $\boldsymbol{q}$, the viable elements with lower indices in $\boldsymbol{q}$ (not of the element values themselves) take priority, such that the notion of greedy selection is coherent. Then, (35a) can be expanded into

$$\hat{\boldsymbol{u}}_{m+1}^{(1)} = \begin{cases} \Big[\boldsymbol{I}_{2Q} - \sum\limits_{q\in odd} \mathrm{diag}(\boldsymbol{e}_q)\Big]\hat{\boldsymbol{u}}_m^{(\eta^*)} & \text{upon confirmation of } \hat{\boldsymbol{k}}^R \text{ from } \hat{\boldsymbol{q}}_m^R, \text{ or} \\ \Big[\boldsymbol{I}_{2Q} - \sum\limits_{q\in even} \mathrm{diag}(\boldsymbol{e}_q)\Big]\hat{\boldsymbol{u}}_m^{(\eta^*)} & \text{upon confirmation of } \hat{\boldsymbol{k}}^I \text{ from } \hat{\boldsymbol{q}}_m^I, \text{ or} \\ \Big[\boldsymbol{I}_{2Q} - \mathrm{diag}(\boldsymbol{e}_{\hat{q}_m})\Big]\hat{\boldsymbol{u}}_m^{(\eta^*)} & \text{otherwise.} \end{cases} \tag{38a}$$

In plain words, (38a) establishes that after the m-th run of the BH-ISTA detector, the initial state of the estimate vector $\hat{\boldsymbol{u}}_{m+1}^{(1)}$ for the next run is either:

a) updated by removing the latest estimate symbol, when neither of $\hat{\boldsymbol{q}}_m^R$ and $\hat{\boldsymbol{q}}_m^I$ can be projected to $\mathcal{K}$, which happens either when the number of indices acquired are insufficient (less than P) to decide on valid estimates of $\boldsymbol{k}^R$ or $\boldsymbol{k}^I$, or when the number of indices are sufficient (P or larger) but none contains valid combinations of indices to any $\boldsymbol{k} \in \mathcal{K}$; or

b) updated by nulling all odd entries of $q \in \{1, 3, \cdots, 2Q-3, 2Q-1\}$, when a hard-decision of $\hat{\boldsymbol{k}}^R$ is confirmed from the projection of $\hat{\boldsymbol{q}}^R$ onto $\mathcal{K}$, which will only happen once throughout the demodulation procedure; or

c) updated by nulling all even entries $q \in \{2, 4, \cdots, 2Q-2, 2Q\}$, when a hard-decision of $\hat{\boldsymbol{k}}_m^I$ is confirmed from the projection of $\hat{\boldsymbol{q}}_m^I$ onto $\mathcal{K}$, which also can happen only once.

Obviously, the only other alternative to those above is when both $\hat{\boldsymbol{k}}^R$ and $\hat{\boldsymbol{k}}^I$ have been acquired, and consequently also the entire set of symbol estimates $\{\hat{s}_1^R, \hat{s}_1^I, \cdots, \hat{s}_P^R, \hat{s}_P^I\}$ have been obtained, in which case the procedure is terminated.

Similarly to the above, the updates of $\boldsymbol{y}_m$ and $\boldsymbol{G}_m$ must also be revised so as to account for the effect of hard-decisions onto $\hat{\boldsymbol{k}}^R$ and $\hat{\boldsymbol{k}}^I$, so as to cancel the effect of hard-decided indices and symbols, and to nullify the channel corresponding to confirmed indices, yielding respectively

$$\boldsymbol{y}_{m+1} = \begin{cases} \boldsymbol{y} - \boldsymbol{G} \cdot \sum\limits_{q\in[2\hat{\boldsymbol{k}}^R-1, \hat{\boldsymbol{q}}_m^I]} \boldsymbol{e}_q\hat{s}_q & \text{upon confirmation of } \hat{\boldsymbol{k}}^R \text{ from } \hat{\boldsymbol{q}}_m^R, \text{ or} \\ \boldsymbol{y} - \boldsymbol{G} \cdot \sum\limits_{q\in[\hat{\boldsymbol{q}}_m^R, 2\hat{\boldsymbol{k}}^I]} \boldsymbol{e}_q\hat{s}_q & \text{upon confirmation of } \hat{\boldsymbol{k}}^I \text{ from } \hat{\boldsymbol{q}}_m^I, \text{ or} \\ \boldsymbol{y}_m - \boldsymbol{G}_m \cdot \boldsymbol{e}_{\hat{q}_m}\hat{s}_{\hat{q}_m} & \text{otherwise,} \end{cases} \tag{38b}$$

$$\boldsymbol{G}_{m+1} = \begin{cases} \boldsymbol{G}_m\Big[\boldsymbol{I}_{2Q} - \sum\limits_{q\in odd} \mathrm{diag}(\boldsymbol{e}_q)\Big] & \text{upon confirmation of } \hat{\boldsymbol{k}}^R \text{ from } \hat{\boldsymbol{q}}_m^R, \text{ or} \\ \boldsymbol{G}_m\Big[\boldsymbol{I}_{2Q} - \sum\limits_{q\in even} \mathrm{diag}(\boldsymbol{e}_q)\Big] & \text{upon confirmation of } \hat{\boldsymbol{k}}^I \text{ from } \hat{\boldsymbol{q}}_m^I, \text{ or} \\ \boldsymbol{G}_m\Big[\boldsymbol{I}_{2Q} - \mathrm{diag}(\boldsymbol{e}_{\hat{q}_m})\Big] & \text{otherwise.} \end{cases} \tag{38c}$$

The procedure described by (32) through (34) and (36) through (38) amount to a greedy – *i.e.*, symbol-by-symbol and index set-by-index set – modification of the GB-ISTA detector introduced earlier, for which it is dubbed as the greedy boxed iterative shrinkage thresholding algorithm for QSM demodulation.

Notice that at the end of the process, estimates $\hat{\boldsymbol{k}}^R$ and $\hat{\boldsymbol{k}}^I$ of the selected dispersion matrix index vectors, as well as hard-decision estimates $\hat{\boldsymbol{s}} = [\hat{s}_1^R, \hat{s}_1^I, \cdots, \hat{s}_P^R, \hat{s}_P^I]$ of the modulated symbols are obtained, from which the corresponding encoded bits $\boldsymbol{b} = [\boldsymbol{b}^R, \boldsymbol{b}^I, \boldsymbol{b}^S]$ can be retrieved at a fraction of the complexity of sphere detection or exhaustive maximum likelihood searches. A diagram illustrating the proposed GB-ISTA QSM receiver is offered in Fig. 9 and a summarized in the form of pseudo-code in Algorithm 3.

C. Complexity Analysis: GB-ISTA versus SotA Receivers

Given that the focus of the OS-QSM operation is with the scalability of the system, considered systems consist of relatively large numbers of transmit antennas (*i.e.*, $n_T \geq 6$) and increasing numbers of transmission slots (*i.e.*, $T \geq 2$), with the digitally-modulated transmit symbols P and the cardinality of corresponding constellation M. To the best of our knowledge, considerations and simulated results on quadrature *spatial-temporal* modulation schemes with such

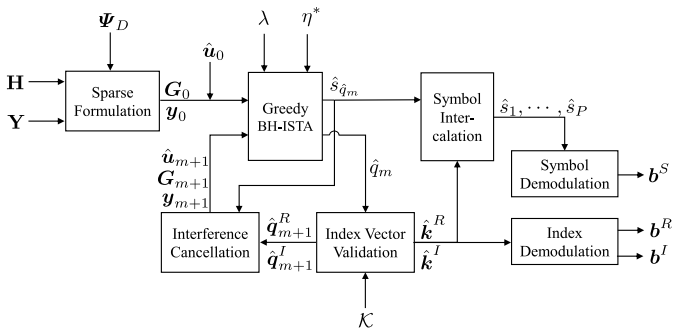


Fig. 9. Schematic diagram depicting the structure of the proposed GB-ISTA receiver for QSM demodulation.

Algorithm 3 Proposed GB-ISTA Receiver for QSM Schemes

Global Quantities: Real-valued projected symbol constellation $\mathcal{S}_R$, set of index vectors $\mathcal{K}$ and threshold factor λ;

Inputs: Received signal $\boldsymbol{y}$ and effective channel matrix $\boldsymbol{G}$.

Outputs: Estimated index and symbol vectors $\hat{\boldsymbol{k}}^R$, $\hat{\boldsymbol{k}}^I$ and $\hat{\boldsymbol{s}}$.

Set $m = 1$ and $\alpha = \text{maxeig}(\boldsymbol{G}^{\mathsf{T}}\boldsymbol{G})$;
Initialize $\boldsymbol{y}_m = \boldsymbol{y}$, $\boldsymbol{G}_m = \boldsymbol{G}$ and $\hat{\boldsymbol{u}}^{(1)}_m = \boldsymbol{0}_{2Q}$;
while $\mathcal{P}_{\mathcal{K}}(\frac{1}{2}[\hat{\boldsymbol{q}}^R_m + 1]) = \emptyset$ or $\mathcal{P}_{\mathcal{K}}(\frac{1}{2}\hat{\boldsymbol{q}}^I_m) = \emptyset$ **do**
  Iterate BH-ISTA (32) until convergence for $\hat{\boldsymbol{u}}^{(\eta^*)}_m$;
  Obtain symbol soft estimate $\tilde{s}_{\hat{q}_m}$ and index hard estimate $\hat{q}_m$ via (33);
  Obtain hard symbol estimate $\hat{s}_{\hat{q}_m}$ via (34);
  Insert index estimate $\hat{q}_m$ into $\hat{\boldsymbol{q}}^R_m$ or $\hat{\boldsymbol{q}}^I_m$ as per (36);
  Construct $\hat{\boldsymbol{u}}^{(1)}_{m+1}$, $\boldsymbol{y}_{m+1}$ and $\boldsymbol{G}_{m+1}$ via (38);
  Increment m by 1;
end while
Output the estimate index vectors as $\hat{\boldsymbol{k}}^R \leftarrow \mathcal{P}_{\mathcal{K}}(\frac{1}{2}[\hat{\boldsymbol{q}}^R_m+1])$ and $\hat{\boldsymbol{k}}^I \leftarrow \mathcal{P}_{\mathcal{K}}(\frac{1}{2}\hat{\boldsymbol{q}}^I_m)$, respectively,
and the estimate symbol vector $\hat{\boldsymbol{s}}$ as the intercalation of $\{\hat{s}_{2\hat{\boldsymbol{k}}^R-1}\}$ and $\{\hat{s}_{2\hat{\boldsymbol{k}}^I}\}$.

increasing parameters have not appeared so far in the literature, due to the prohibitive computational complexity of existing ML-based receivers.

In light of the latter remark, let us start by assessing the decoding complexity of scaled QSM systems, in particular by deriving the complexity orders of the conventional ML and SCMB-SD approaches, and of the proposed GB-ISTA algorithm described in Section IV.

For any given n_T, T and P, the brute-force ML decoder requires a search among all possible $\left\lfloor \binom{Tn_T}{P} \right\rfloor_{2^\times}$ antenna activation patterns, independently selected according to $\{\boldsymbol{k}^R, \boldsymbol{k}^I\} \in \mathcal{K}$ to transmit the real and imaginary parts of the P digitally modulated symbols in $\boldsymbol{s} \in \mathcal{S}^P$, as well as another search, for each possible activation pattern, of all possible P-tuples of symbols selected from the constellation $\mathcal{S}$ of cardinality M. Assuming, idealistically and for simplicity, that each search consumes a single floating point operation (flop), the ML search process alone yields a complexity order lower-bounded by $\mathcal{O}\big(4^{\lfloor \log_2 \binom{Tn_T}{P} \rfloor} \times M^P\big)$, in order to detect $2 \cdot \lfloor \log_2 \binom{Tn_T}{P} \rfloor + P \cdot \log_2 M$ bits, which even for moderately small P and M quickly become unfeasible. For example, a search over 16,777,216 combinations is required to detect the 24 bits of each transmit signal in a relatively small system with $n_T = 6$, $T = 3$, $P = 3$ and $M = 4$. Just doubling the number of transmit antennas to $n_T = 12$, with other parameters unchanged, the complexity of the ML search space already surges to 10^9 combinations, for a mild increase to 30 bits per transmission, while keeping $n_T = 6$ and doubling the number of transmit symbols to $P = 6$ requires a search over more than 10^{12} combinations in order to detect only 40 bits.

Taking the most significant operations required to perform each ML search into account,[9] the order of complexity of the ML receiver to decode each bit of QSM schemes becomes

$$\mathcal{O}\big(\overbrace{12T^3 n_T n_R}^{\text{construction of } \boldsymbol{\Phi}_H \cdot \boldsymbol{\Xi}_A \text{ and } \boldsymbol{\Phi}_H \cdot \boldsymbol{\Xi}_B \text{ as in (23)}} + (4Tn_R(2P+1) - 1) \cdot 4^{\lfloor \log_2 \binom{Tn_T}{P} \rfloor} \cdot M^P\big) > \mathcal{O}\big(n_R \cdot P \cdot T^{P+1} \cdot M^P \cdot n_T^{P-1}\big), \tag{39}$$

with the lower bound obtained by keeping only the higher-order terms and neglecting coefficients.

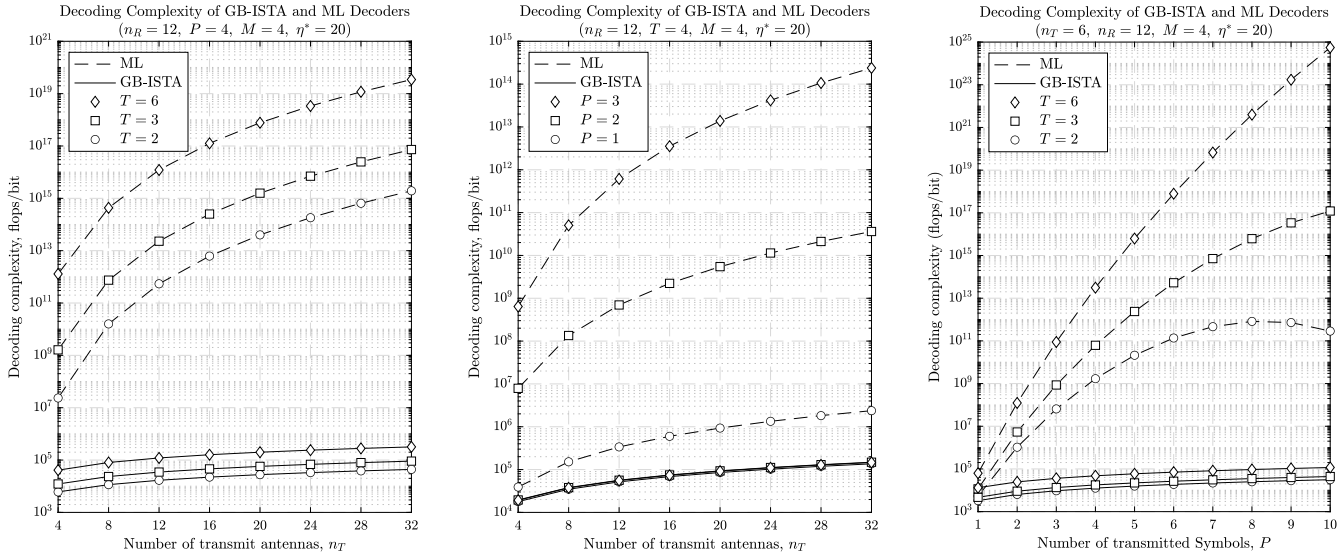
The practical unfeasibility of ML-based detection of QSM systems is clearly highlighted by (39), as it exposes the fact that the number of transmit symbols P is a complexity order exponent of all theoretically scalable quantities n_T, T and M. Next, let us show that this challenge cannot be satisfactorily mitigated by the SD approach [30], [31]. To that end, consider again idealistically and for simplicity, that SD can reduce the search radius to a single symbol, such that the factor M^P in (39) can be neglected. In other words, we find that the order of complexity associated to SD-based QSM receivers can, at best, be reduced to

$$\mathcal{O}\big(n_R \cdot P \cdot T^{P+1} \cdot n_T^{P-1}\big). \tag{40}$$

From the latter it can be concluded that, in the context of scalable QSM schemes, the only advantage of SD is to enable scaling of the digital constellation cardinality M, which not only impacts negatively on the corresponding BER, but also is not the most significant factor in increasing the SE of the system, since the total number of bits conveyed by a QSM scheme is $B = 2 \cdot \lfloor \log_2 \binom{Tn_T}{P} \rfloor + P \cdot \log_2 M$, such that even for mildly large n_T, T and P we have $2 \cdot \lfloor \log_2 \binom{Tn_T}{P} \rfloor \gg P \cdot \log_2 M$. In summary, it can be concluded that sphere detection is not particularly useful as an enabler of spectrally-efficient scalable QSM, from a receiver perspective.

Finally, let us address the computation complexity of the proposed GB-ISTA. For starters, observe from (32) through (37) that the GB-ISTA receiver obtains the spatially encoded bits $\boldsymbol{b}^R$ and $\boldsymbol{b}^I$ not from a search, but directly from the sparse-recovery process, *i.e.*, the value and locations of non-zero elements of $\hat{\boldsymbol{u}}$. As a consequence of removing such combinatorial search, the impact of the scalable parameters n_T, T and P onto GB-ISTA is significantly smaller, as demonstrated by the following complexity analysis of the steps of Algorithm 3.

[9]We omit details due to page limitations and since the ML receiver is well-known.



(a) Fixed P, various T, against n_T. (b) Fixed T, various P, against n_T. (c) Fixed n_T, various T, against P.

Fig. 10. Effect of scalable parameters on the complexity of QSM receivers.

1) Algorithm 3 takes as input the effective matrix $\boldsymbol{G}$ given in (28), whose construction requires evaluating the product of the sparse block-diagonal matrix $\boldsymbol{\Phi}_H \in \mathbb{R}^{2Tn_R \times 2Tn_T}$, which contains $2n_T$ non-zero entries per row, against the matrix $\boldsymbol{\Psi}_D \in \mathbb{R}^{2Tn_T \times 2Tn_T}$, which has T non-zero entries per column, yielding a cost of $2T(2Tn_R)(2Tn_T) = 8T^3 n_T n_R$ flops, since typically we have (in scaled QSM schemes) $T \ll 2n_T$.
2) Next,[10] the GB-ISTA receiver performs multiple runs of evaluating (32), the first step to which is computing $\hat{\boldsymbol{u}}^{(\eta)} + \frac{1}{\alpha}\boldsymbol{G}^{\mathsf{T}}\big(\boldsymbol{y} - \boldsymbol{G}\,\hat{\boldsymbol{u}}^{(\eta)}\big)$. Naively counted, that operation would cost $(2Tn_T)(2Tn_R) + (2Tn_R) + (2Tn_R)(2Tn_T) + (2Tn_R) = 8T^2 n_T n_R + 4Tn_R$ flops to accomplish, but since the sparsity of $\hat{\boldsymbol{u}}^{(\eta)}$ quickly reduces to the actual value $2P$, as shown in Fig. 8a, the complexity of that step is more precisely estimated at $8PTn_R + 4Tn_R = 4Tn_R(2P+1)$ flops. Then, including $2Tn_T$ flops required for the Boxed-Hard thresholding function Π, and remembering η^* iterations are necessary, the total cost associated with each evaluation run of (32) can be estimated at $\eta^*(4Tn_R(2P+1) + 2Tn_T)$ flops.
3) After convergence of (32), the receiver obtains the sparse estimate vector $\hat{\boldsymbol{u}}_m^{(\eta^*)}$, from which the soft symbol estimate $\tilde{s}_{\hat{q}_m}$ is extracted at negligible cost via maximum value search [49], along with the estimate index $\hat{q}_m$ given by the position of $\tilde{s}_{\hat{q}_m}$ in $\hat{\boldsymbol{u}}_m^{(\eta^*)}$, as expressed in (33). With these quantities at hand, up to $\sqrt{M}$ flops are consumed to obtain the hard symbol estimate $\hat{s}_{\hat{q}_m}$ as per (34).
4) Considering that the cost of the element, interference and column removals expressed by (38a) through (38c) are negligible, the next significant cost of the receiver is the validation of acquired indices. In particular, after at least P runs, when a sufficient number of position indices $\hat{q}_m$ have been detected to construct any or both of the index subsequences $\hat{\boldsymbol{q}}_m^R$ and/or $\hat{\boldsymbol{q}}_m^I$, and map them to corresponding estimate index vectors $\hat{\boldsymbol{k}}^R$ and/or $\hat{\boldsymbol{k}}^I$ as per (37), said estimates need be validated against the optimal set of index vectors $\mathcal{K}$. Assuming the cost of such operation is of order $\mathcal{O}(1)$, this step contributes to the total complexity of the GB-ISTA detector with an additional cost of P flops.
5) Lastly, as described in line 11 of Algorithm 3, the GB-ISTA outputs both the pair of estimate index vectors $\hat{\boldsymbol{k}}^R$ and $\hat{\boldsymbol{k}}^I$, as well as the digitally modulated symbol vector estimate $\hat{\boldsymbol{s}} \in \mathcal{S}^P$, which requires the intercalation of the real and imaginary parts, at estimated cost of P flops.

From the above, the total complexity order of the GB-ISTA can be estimated at

$$\mathcal{O}\Big(\overbrace{8T^3 n_T n_R}^{\text{construction of } \boldsymbol{G}} + \underbrace{2P}_{\text{number of runs}} \big(\overbrace{\eta^*(4Tn_R(2P+1) + 2Tn_T)}^{\text{evaluation of (32)}} + \overbrace{\underbrace{\sqrt{M} + \tfrac{1}{2}}_{\text{evaluation of (36) and validation of } \hat{\boldsymbol{q}}_m^I \text{ or } \hat{\boldsymbol{q}}_m^R}}^{\text{intercalation of } \{\hat{s}_{2\hat{\boldsymbol{k}}^R-1}\} \text{ and } \{\hat{s}_{2\hat{\boldsymbol{k}}^I}\} \text{ into } \hat{\boldsymbol{s}}} + \overbrace{P}^{} \big) \Big). \tag{41}$$

The per-bit complexity orders of the ML and the proposed GB-ISTA decoders, obtained by dividing the expressions in (39) and (41) by the number of bits per transmission $B = 2\lfloor \log_2 \binom{Tn_T}{P} \rfloor + P \log_2 M$, are compared in Fig. 10 for various settings in terms of the scalable parameters n_T, T and P.

V. Performance Analysis and Discussion

Empowered by the significant reduction in complexity obtained by GB-ISTA over ML detection, as shown above, we proceed to assess the BER performance of the proposed

[10] We ignore the cost of computing α as per line 1 of Algorithm 3, under the argument that for large systems, the largest eigenvalues $\boldsymbol{G}^{\mathsf{T}}\boldsymbol{G}$ converges almost sure to a constant dependent only on the structure and energy of $\boldsymbol{G}$.

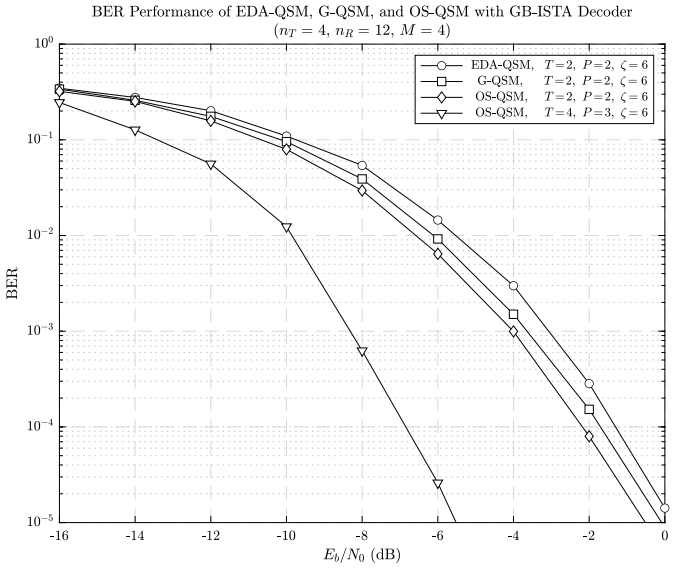


Fig. 11. Effect of the proposed contributions on GB-ISTA-detected BER performance in a given fixed system.

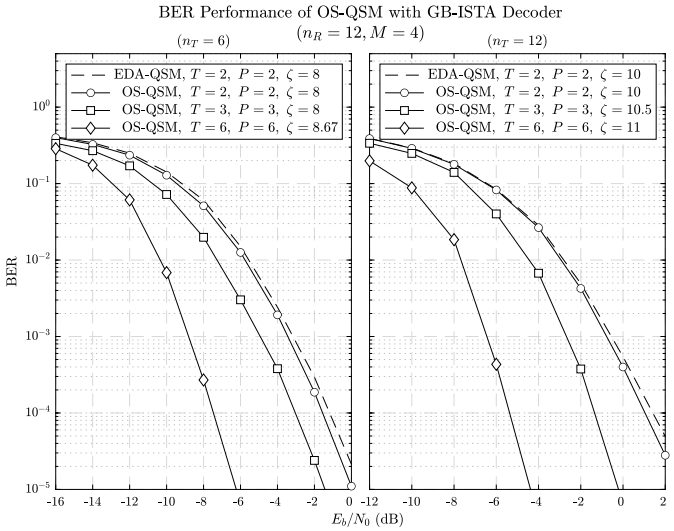


Fig. 12. Effect of temporal scalability on BER performance of GB-ISTA-detected OS-QSM schemes.

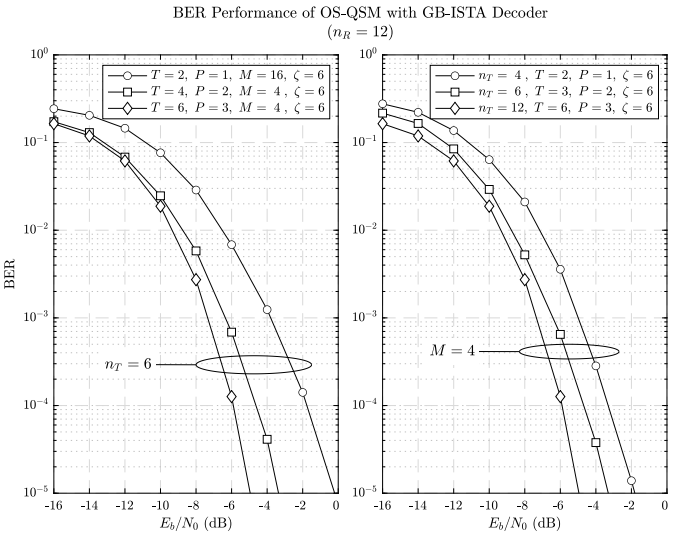


Fig. 13. Effect of scalability on BER performance of GB-ISTA-detected OS-QSM schemes with fixed SE.

OS-QSM scheme, decoded via the GB-ISTA. In general, our simulated experiments aim to further demonstrate that the proposed OS-QSM are feasible with relatively large numbers of transmit antennas, and can achieve very low BER at very low E_b/N_0 with rather high spectral efficiencies, whilst using relatively few spatial-temporal resources per transmission.

In Fig. 11, the clear advantageous effect of our contributions from Subsection III is illustrated. Namely, notice progressively from the SotA EDA-QSM scheme (*circle*), the aggregated improvements via: a) the addition of coding gain optimality (*square*), b) the addition of optimal resource allocation (*diamond*), and c) the addition of temporal scalability (*triangle*).

Then in Fig. 12, the effect of temporal scalability in the proposed OS-QSM is illustrated. The system in each subplot is fixed on n_T and P/T for a fair comparison, and hence the different plots compare the improvement gained when increasing T to a larger value available. Two significant improvements may be observed, which are a) clear BER performance gain, and b) SE gain.

Fig. 13 compares the BER performances of the proposed method for various values of n_T, T and P, with M adjusted such that all curves corresponds to systems with the *same spectral efficiency*. Another important fact that should be highlighted from the figures, is that the results shown are actually simulated, down to rather low BERs, using a usual computer ($i.e.$ no particularly powerful machine was required), and for settings which are virtually impossible to simulate with ML- or SD-based receivers. The latter point is strengthened by the results of the right side of Fig. 13, which includes a curve for a system with $n_T = 12$, which serves to further highlight the true feasibility of the proposed GB-ISTA receiver.

One criticism that could be raised about the results so far is, however, that values of P adopted thereby are not the optimal ones for the corresponding T and n_T, as per (22). We once again clarify that the parameterization used in Fig. 13 is such that all systems have the same SE, so as to allow their direct comparison under equivalent conditions, which is, incidentally also the reason why all curves are plotted against E_b/N_0 as opposed to SNR. In any case, in order to dispel any doubts about the ability of the proposed OS-QSM design and of the corresponding GB-ISTA receiver to actually achieve *feasible* and optimized spectral efficiency combined with low BERs, we show in Fig. 14 additional results obtained by varying P up to the optimal value given by (22). We remark that, due to the floor operation in the expression of the achievable SE given by 16, values of P adjacent to that given by (22) – $i.e.$, P^* – may *also* be optimum, and result exactly the *same* SE.

For instance, with $n_T = 6$, $T = 2$ and $M = 4$, as is the case of Fig. 13, (22) yields $P^* = 8$, but the values $P = \{7, 8, 9\}$ all result in $\zeta = 16$ in (16). Similarly, $P = \{10, 11, 12, 13\}$ all yield the largest SE of $\zeta = 17$ with $n_T = 6$, $T = 3$ and $M = 4$ as is the case of Fig. 14.

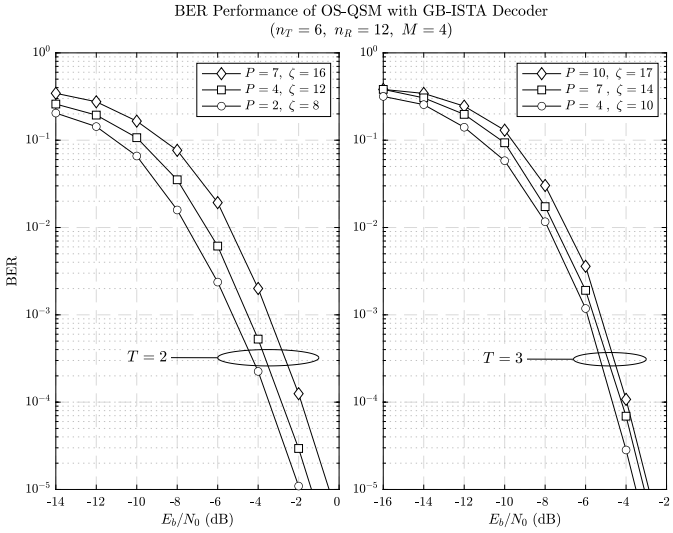


Fig. 14. Effect of scaling P on BER performance of GB-ISTA-detected OS-QSM schemes.

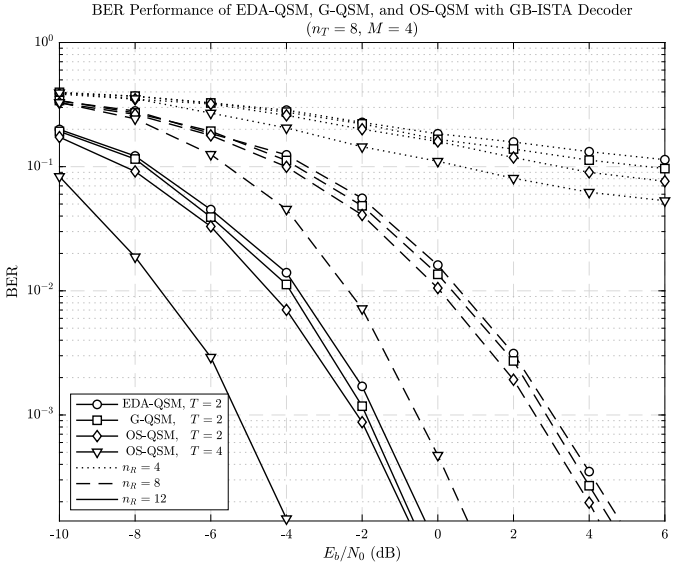


Fig. 15. Comparison of proposed OS-QSM and SotA schemes with variable n_R.

With these remarks made, turning to the results obtained in Fig. 14, it can be seen that only a very mild degradation of BER is observed when up-scaling P, which is in fact smaller in for larger T as shown in Fig. 14, which is a small and fair price to pay for almost doubling spectral efficiency of the system. We clarify that the slight BER degradation observed when up-scaling the ratio P/T towards SE optimality is a consequence of the corresponding reduction of sparsity in the vectorized received signal, which tends to be less critical in systems with higher diversity and coding gains, as a result of up-scaling n_T, T or both. That trend is in fact observable in Fig. 14, as the gap between BER curves narrows as $T = 2$ is increased to $T = 3$.

Lastly, Fig. 15 offers a comparison between the proposed OS-QSM and SotA methods with $n_R = 4, 8$ and 12. Although the performance of all methods degrades significantly for $n_T > n_R$, as a result of overloading, it can be seen that the proposed scheme maintains a consistent and significant BER advantage over the existing alternatives in all cases.

VI. Conclusion

We proposed new transmitter and receiver designs for QSM schemes, focusing on their scalability in terms of the number of transmit antennas n_T, of transmission instances T and of encoded M-ary symbols P, as well as on their performance optimization in terms of SE, diversity and coding gains. The contributions are motivated by the demonstrated fact that, in order for SE optimality to be achieved, QSM schemes must scale n_T, T and P, which is not possible in SotA methods such as those in [7], [23]–[25]. At the transmitter, the proposed OS-QSM scheme differs from SotA alternatives in that its dispersion matrices are designed based on the $T \times T$ FDFR STBCs of [33], [34], and in that dispersion matrix index selection is performed via a new greedy algorithm, which ensures that all spatial-temporal resources of the transmitter are utilized evenly. In turn, at the receiver, the proposed art contributes with a new ISTA-based receiver, which thanks to its reliance on the sparse structure of QSM signaling, eliminates the combinatorial nature of existing ML- or SD-based approaches, further enabling the scaling of the system from a feasibility perspective. In fact, a complexity analysis is offered, which shows that the proposed GB-ISTA receiver enjoys a complexity order that is cubic on T, quadratic on P, and only linear on n_T, in contrast to the ML and SD detectors which have geometric complexities on T and n_T, with P as exponent, rendering them unfeasible in the scaled scenario. Simulation results for set-ups of scales never before shown in related literature, corroborate both the high performance and feasibility of the proposed OS-QSM scheme and GB-ISTA receiver.

References

[1] R. Y. Mesleh, H. Haas, S. Sinanovic, C. W. Ahn, and S. Yun, "Spatial modulation," *IEEE Trans. Veh. Technol.*, vol. 57, no. 4, pp. 2228–2241, Jul. 2008.

[2] R. Mesleh, H. Haas, C. W. Ahn, and S. Yun, "Spatial modulation—A new low complexity spectral efficiency enhancing technique," in *Proc. 1st Int. Conf. Commun. Netw. China*, 2006, pp. 1–5.

[3] J. Jeganathan, A. Ghrayeb, L. Szczecinski, and A. Ceron, "Space shift keying modulation for MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3692–3703, Jul. 2009.

[4] T. S. Rappaport, Y. Xing, G. R. Maccartney, A. F. Molisch, E. Mellios, and J. Zhang, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks—With a focus on propagation models," *IEEE Trans. Antennas Propag.*, vol. 65, no. 12, pp. 6213–6230, Dec. 2017.

[5] T. S. Rappaport *et al.*, "Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond," *IEEE Access*, vol. 7, pp. 78729–78757, 2019.

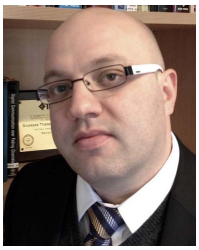
[6] A. Younis, N. Serafimovski, R. Mesleh, and H. Haas, "Generalised spatial modulation," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Nov. 2010, pp. 1498–1502.

[7] R. Mesleh, S. S. Ikki, and H. M. Aggoune, "Quadrature spatial modulation," *IEEE Trans. Veh. Technol.*, vol. 64, no. 6, pp. 2738–2742, Jun. 2015.

[8] A. M. Abu-Hudrouss, M.-T.-O. E. Astal, A. H. Al Habbash, and S. Aissa, "Signed quadrature spatial modulation for MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 2740–2746, Mar. 2020.

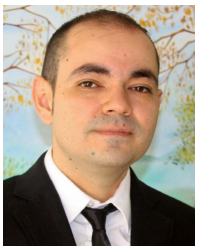
[9] A. Habbash *et al.*, "Extended signed quadrature spatial modulation system with multi-user support," *IEEE Trans. Broadcast.*, vol. 67, no. 4, pp. 1–11, Apr. 2021.
[10] B. Vo and H. H. Nguyen, "Improved quadrature spatial modulation," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2017, pp. 1–5.
[11] Y. Celik, "Fully improved quadrature spatial modulation," *Arabian J. Sci. Eng.*, vol. 46, no. 10, pp. 9639–9647, Feb. 2021.
[12] A. B. Saleem, S. A. Hassan, H. Jung, S. Garg, G. Kaddoum, and M. Guizani, "Full-duplex quadrature spatial modulation for multi-antenna systems," *IEEE Netw.*, vol. 35, no. 5, pp. 226–233, Sep. 2021.
[13] M. Di Renzo and H. Haas, "Bit error probability of SM-MIMO over generalized fading channels," *IEEE Trans. Veh. Technol.*, vol. 61, no. 3, pp. 1124–1144, Mar. 2012.
[14] S. Sugiura, S. Chen, and L. Hanzo, "Coherent and differential space-time shift keying: A dispersion matrix approach," *IEEE Trans. Commun.*, vol. 58, no. 11, pp. 3219–3230, Nov. 2010.
[15] S. Sugiura, S. Chen, and L. Hanzo, "Generalized space-time shift keying designed for flexible diversity-, multiplexing- and complexity-tradeoffs," *IEEE Trans. Wireless Commun.*, vol. 10, no. 4, pp. 1144–1153, Apr. 2011.
[16] S. Sugiura, S. Chen, and L. Hanzo, "A universal space-time architecture for multiple-antenna aided systems," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 2, pp. 401–420, 2nd Quart., 2012.
[17] E. Başar, U. Aygolu, E. Panayirci, and H. V. Poor, "Space-time block coded spatial modulation," *IEEE Trans. Commun.*, vol. 59, no. 3, pp. 823–832, Mar. 2011.
[18] M.-T. Le, V.-D. Ngo, H.-A. Mai, X. N. Tran, and M. Di Renzo, "Spatially modulated orthogonal space-time block codes with non-vanishing determinants," *IEEE Trans. Commun.*, vol. 62, no. 1, pp. 85–99, Jan. 2014.
[19] L. Wang and Z. Chen, "Spatially modulated diagonal space time codes," *IEEE Commun. Lett.*, vol. 19, no. 7, pp. 1245–1248, Jul. 2015.
[20] M.-T. Le, T.-D. Nguyen, X.-N. Tran, and V.-D. Ngo, "On the combination of double space time transmit diversity with spatial modulation," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 170–181, Jan. 2018.
[21] X. Li and L. Wang, "High rate space-time block coded spatial modulation with cyclic structure," *IEEE Commun. Lett.*, vol. 18, no. 4, pp. 532–535, Feb. 2014.
[22] G. Huang, C. Li, S. Aissa, and M. Xia, "Parallel quadrature spatial modulation for massive MIMO systems with ICI avoidance," *IEEE Access*, vol. 7, pp. 154750–154760, 2019.
[23] L. Wang, Z. Chen, Z. Gong, and M. Wu, "Diversity-achieving quadrature spatial modulation," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 10764–10775, Dec. 2017.
[24] L. Wang and Z. Chen, "Stacked Alamouti based spatial modulation," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 336–349, Sep. 2019.
[25] L. Wang and Z. Chen, "Enhanced diversity-achieving quadrature spatial modulation with fast decodability," *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6165–6177, Jun. 2020.
[26] S. Sezginer, H. Sari, and E. Biglieri, "On high-rate full-diversity 2×2 space-time codes with low-complexity optimum detection," *IEEE Trans. Commun.*, vol. 57, no. 5, pp. 1532–1541, May 2009.
[27] S. Guo, H. Zhang, P. Zhang, S. Dang, C. Liang, and M.-S. Alouini, "Signal shaping for generalized spatial modulation and generalized quadrature spatial modulation," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4047–4059, Aug. 2019.
[28] R. Mesleh, S. Althunibat, and A. Younis, "Differential quadrature spatial modulation," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3810–3817, Sep. 2017.
[29] J. Jaldén and B. Ottersten, "On the complexity of sphere decoding in digital communications," *IEEE Trans. Signal Process.*, vol. 53, no. 4, pp. 1474–1484, Apr. 2005.
[30] B. Hassibi and H. Vikalo, "On the sphere-decoding algorithm I. Expected complexity," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2806–2818, Aug. 2005.
[31] D. Seethaler, J. Jaldén, C. Studer, and H. Bölcskei, "On the complexity distribution of sphere decoding," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 5754–5768, Sep. 2011.
[32] J. Belfiore, G. Rekaya, and E. Viterbo, "The golden code: A 2×2 full-rate space-time code with nonvanishing determinants," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1432–1436, Apr. 2005.
[33] F. Oggier, G. Rekaya, J.-C. Belfiore, and E. Viterbo, "Perfect space-time block codes," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3885–3902, Sep. 2006.
[34] P. Elia, K. Kumar, S. Pawar, P. Kumar, and H.-F. Lu, "Explicit space–time codes achieving the diversity–multiplexing gain tradeoff," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3869–3884, Sep. 2006.
[35] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1414–1457, Nov. 2004.
[36] E. C. Marques, N. Maciel, L. Naviner, H. Cai, and J. Yang, "A review of sparse recovery algorithms," *IEEE Access*, vol. 7, pp. 1300–1322, 2019.
[37] P. Stanica, "Good lower and upper bounds on binomial coefficients," *J. Inequal. Pure Appl. Math.*, vol. 2, no. 3, p. 30, 2001.
[38] E. Castaneda, A. Silva, A. Gameiro, and M. Kountouris, "An overview on resource allocation techniques for multi-user MIMO systems," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 239–284, 1st Quart., 2017.
[39] S. Pemmaraju and S. S. Skiena, *Computational Discrete Mathematics: Combinatorics and Graph Theory With Mathematica*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[40] Y. C. Eldar, *Compressed Sensing: Theory Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
[41] A. Aïssa-El-Bey, D. Pastor, S. M. A. Sbaï, and Y. Fadlallah, "Sparsity-based recovery of finite alphabet solutions to underdetermined linear systems," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 2008–2018, Apr. 2015.
[42] R. A. Stoica *et al.*, "Frame theory and fractional programming for sparse recovery-based mmwave channel estimation," *IEEE Access*, vol. 7, pp. 150757–150774, 2019.
[43] H. Iimori *et al.*, "Robust symbol detection in large-scale overloaded NOMA systems," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 512–533, 2021.
[44] L. Liu, Y. Li, C. Huang, C. Yuen, and Y. L. Guan, "A new insight into GAMP and AMP," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8264–8269, Aug. 2019.
[45] L. Liu, C. Yuen, Y. L. Guan, Y. Li, and C. Huang, "Gaussian message passing for overloaded massive MIMO-NOMA," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 210–226, Jan. 2019.
[46] K. H. Ngo *et al.*, "Multi-user detection based on expectation propagation for the non-coherent SIMO multiple access channel," *IEEE Trans. Wireless Commun.*, vol. 19, no. 9, pp. 6145–6161, Sep. 2020.
[47] J. Barbier, M. Dia, and N. Macris, "Universal sparse superposition codes with spatial coupling and GAMP decoding," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5618–5642, Dec. 2019.
[48] R. Hayakawa and K. Hayashi, "Discreteness-aware approximate message passing for discrete-valued vector reconstruction," *IEEE Trans. Signal Process.*, vol. 66, no. 24, pp. 6443–6457, Dec. 2018.
[49] H. S. Wilf, *Algorithms Complexity*. Boca Raton, FL, USA: CRC Press, 2020.

Hyeon Seok Rou (Graduate Student Member, IEEE) received the B.Sc. degree in electrical and computer engineering from Jacobs University Bremen, Germany, in 2021, where he is currently pursuing the Ph.D. degree in electrical engineering, funded by a research project from the Wireless Communications Technologies Group, Continental A.G. His research interests include joint communications and sensing (JCAS), Bayesian statistics, and mmWave/sub-THz MIMO wireless communications.



Giuseppe Thadeu Freitas de Abreu (Senior Member, IEEE) received the B.Eng. degree in electrical engineering and a specialization (Latu Sensu) degree in telecommunications engineering from the Universidade Federal da Bahia (UFBa), Salvador, Bahia, Brazil, in 1996 and 1997, respectively, and the M.Eng. and D.Eng. degrees in physics, electrical and computer engineering from Yokohama National University, Japan, in March 2001 and March 2004, respectively. He was a Post-Doctoral Fellow and later an Adjunct Professor (Docent) of statistical signal processing and communications theory at the Department of Electrical and Information Engineering, University of Oulu, Finland, from 2004 to 2006 and from 2006 to 2011, respectively. Since 2011, he has been a Professor of electrical engineering at Jacobs University Bremen, Germany. From April 2015 to August 2018, he was a Full Professor at the Department of Computer and Electrical Engineering, Ritsumeikan University, Japan. His research interests include communications and signal processing, including communications theory, estimation theory, statistical modeling, wireless localization, cognitive radio, wireless security, MIMO systems, ultrawideband and millimeter wave communications, full-duplex and cognitive radio, compressive sensing, energy harvesting networks, random networks, connected vehicles networks, joint communications and sensing, and many others. He was a recipient of the Uenohara Award by Tokyo University in 2000 for his master's thesis work. He was a co-recipient of best paper awards at several international conferences, and was awarded the JSPS, Heiwa Nakajima, and NICT Fellowships (twice) in 2010, 2013, 2015, and 2018, respectively. He served as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2009 to 2014 and the IEEE TRANSACTIONS ON COMMUNICATIONS from 2014 to 2017, and as an Executive Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2018 to 2021.

Hiroki Iimori (Graduate Student Member, IEEE) received the B.Eng. and M.Eng. degrees (Hons.) in electrical and electronic engineering from Ritsumeikan University, Kyoto, Japan, in 2017 and 2019, respectively, and the Ph.D. degree (Hons.) in electrical engineering from Jacobs University Bremen, Germany, in 2022. He was a Visiting Scholar at the University of Toronto, Canada, in 2020. In 2021, he was a Research Intern with Ericsson Radio S&R Research Laboratory, Yokohama, Japan. His research interests include optimization theory, wireless communications, and signal processing. He was awarded the YKK Doctoral Fellowship by Yoshida Scholarship Foundation, the IEICE Young Researcher of the Year Award by the IEICE Smart Radio Committee in 2020, amongst many others.



David González G. (Senior Member, IEEE) received the master's degree in mobile communications and the Ph.D. degree in signal theory and communications from the Universitat Politécnica de Catalunya, Spain. He worked as a Post-Doctoral Fellow at Aalto University, Finland (2014–2017). He also worked at Panasonic Research and Development Center, Germany, and since 2018, he has been with Continental AG, Germany. His research is focused on diverse aspects of cellular networks and wireless communications, including interference modeling, radio access modeling and optimization, and vehicular communications (V2X). Since 2017, he has contributed to the 3GPP for 5G standardization.

Osvaldo Gonsa received the Ph.D. degree in electrical and computer engineering from Yokohama National University, Japan, in 1999, and the M.B.A. degree from the Kempten School of Business, Germany, in 2012. He worked in research and standardization in the areas of core and radio access network. He is currently the Head of the Wireless Communications Technologies Group, Continental AG, Frankfurt, Germany. Since 2015, he has been an Advisor of the German Federal Ministry of Economy and Energy for the "PAiCE" projects. He also serves as a member for the GSMA Advisory Board for automotive and the 6GKom Project of the German Federal Ministry of Education and Research.