IEEE Open Journal of the Communications Society

Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version 11 January, 2024.

Digital Object Identifier 10.1109/OJCOMS.2024.011100

Constrained Symbol-Level Noise-Guessing Decoding with Antenna Sorting for Massive MIMO

SAHAR ALLAHKARAM[1,2] *(Student Member, IEEE)*, FRANCISCO A. MONTEIRO [1,2] *(Member, IEEE)*, AND IOANNIS CHATZIGEORGIOU[3] *(Senior Member, IEEE)*

[1]Instituto de Telecomunicações, Lisbon, Portugal
[2]ISCTE – Instituto Universitário de Lisboa, Portugal
[3]School of Computing & Communications, Lancaster University, UK

CORRESPONDING AUTHOR: Francisco A. Monteiro (e-mail: francisco.monteiro@lx.it.pt)

This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., and, when eligible, co-funded by EU funds under project/support UID/50008/2025 – Instituto de Telecomunicações. Sahar Allahkaram is funded by a merit scholarship from ISCTE – Instituto Universitário de Lisboa.

ABSTRACT Supporting ultra-reliable and low-latency communication (URLLC) is a challenge in current wireless systems. Channel codes that generate large codewords improve reliability but necessitate the use of interleavers, which introduce undesirable latency. Only short codewords can eliminate the requirement for interleaving and reduce decoding latency. This paper suggests a coding and decoding method which, when combined with the high spectral efficiency of spatial multiplexing, can provide URLLC over a fading channel. Random linear coding and modulation are used to transmit information over a massive multiple-input multiple-output (mMIMO) uplink channel, followed by zero-forcing detection and guessing random additive noise decoding (GRAND) at a receiver. This work considers symbol-level GRAND, which is a variant of GRAND that was originally proposed for single-antenna systems employing square M-ary quadrature amplitude modulation, and generalizes it to schemes that combine spatial multiplexing with any M-ary modulation method. The paper studies the impact of the orthogonality defect of the underlying mMIMO lattice on symbol-level GRAND, and proposes to leverage side-information that comes from the mMIMO channel-state information and relates to the reliability of each receive antenna. Additionally, a lightweight membership test is introduced to reduce the number of error patterns that undergo full membership tests, by making use of a row in the parity-check matrix that eliminates candidate error patterns. All proposals reduce the decoding speed without compromising the decoding performance. The proposed decoder operating at the symbol level, when combined with antenna sorting and syndrome-constrained decoding, has the potential to reduce complexity by 90% when compared to bit-level GRAND in some of the tested configurations.

INDEX TERMS Antenna sorting, constrained decoding, guessing random additive noise decoding (GRAND), massive multiple-input multiple-output (mMIMO), random linear codes (RLCs), symbol-level decoding, ultra-reliable and low-latency communications (URLLC)

I. Introduction

In addition to other crucial requirements for the sixth generation (6G) of wireless networks, such as low energy consumption, high scalability, stability, security, and ubiquitous connectivity, the physical layer of wireless communications will have to significantly contribute to the goal of ultra-reliable and low-latency communications (URLLC). To meet the important requirements of applications like the industrial internet of things (IIoT), virtual reality, or self-driving cars, URLLC's main objectives are to reduce latency to 1 ms while concurrently guaranteeing at least 99.999% dependability [1]. Using error-correcting codes with short codewords is one way of achieving the sought low-latency objective, because that allows to discard the interleavers that are

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

typically employed in wireless links to make the errors look independent and identically distributed (i.i.d.) [2]. However, developing codes with large codewords was prioritized in pre-5G systems to reach Shannon's capacity [3], [4]. An interest in codes from the 1960s, such as Reed-Solomon and BCH codes, was rekindled, aiming at URLLC applications [5]. While these codes can have short codewords, they only exist for a limited number of code rates. Contrary to that, random linear codes (RLCs) can be constructed with any code rate, even though decoding long RLCs is impractical [6].

It was previously known that short RLCs could be decoded using trellis decoding [7]–[14] or information set decoders [15], however, given the historical emphasis on capacity-achieving codes (with long codewords), that path of research seems to have been abandoned by the coding community. Recently, noise-guessing decoding has been proposed as a universal decoding technique for codes with moderate length or sufficiently high rate, which are particularly suited for wireless URLLC [16]. The method, known as guessing random additive noise decoding (GRAND) allows maximum likelihood (ML) decoding with a considerably reduced complexity, chiefly because it focuses on "decoding the noise" rather than the codewords, by taking advantage of the entropy of the noise being much lower than the entropy of the codewords. The sole requirement is that a code membership test exists to decide whether some word is a valid codeword. Consequently, GRAND can perform ML decoding of binary or nonbinary linear block codes (such as polar codes [17], [18], BCH codes [18], [19], or Hamming codes) without the need to compute a trellis or store a large table.

Also, GRAND opened doors to using RLCs, known to be capacity-achieving in the asymptotic regime (i.e., with infinite length codewords) in the binary symmetric channel (BSC) [3], [4], and they also reach capacity in the finite-blocklength regime [16], [20], [21], which is the regime of interest for URLLC applications. Several recent research works have shown that RLCs supersede the performance of polar codes of the same length and rate in the classical case [22], [23]. Most importantly, while off-the-shelf nonrandom codes, such as polar codes, do not exist for any desired pair of code length and code rate, one has great flexibility of choice regarding the length and code rate when employing RLCs with GRAND [2], [16], [24]. For higher spectral efficiency, GRAND has been proposed in combination with massive multiple-input multiple-output (mMIMO) in [23]. The ideas behind GRAND have also been adapted to allow the decoding of quantum random linear codes in a practical manner [25], [26], and also to decode quantum stabilizer codes with a given structure (i.e., nonrandom known codes) [27].

Several works have recently proposed enhancements to reduce the complexity of decoders based on guessing techniques. Guessing codeword decoding (GCD) is an ML decoding algorithm [28] that orders error patterns based on a newly defined metric that serves as a "soft Hamming weight" for each candidate error pattern, and that allows them to be sorted in decreasing probability; moreover, the decoder only tests error patterns that have a given structure associated with the systematic form of the code (which is always available via Gaussian elimination). The ordered reliability direct error pattern testing (ORDEPT) decoder [29] introduced a technique to test error patterns that are at a Hamming distance-one of a currently tested error pattern. This early tests will change the optimal order of testing error patterns and may lead to finding a codewords that is not the ML solution. To compensate, the authors proposed populating a list of error candidates that lead to valid codewords and only make a decision in the end. In doing that, ORDEPT attains ML performance, and outperforms both the (sub-optimal) ordered reliability bits GRAND (ORBGRAND) [18] and the even more complex first soft GRAND decoder [30]. Soft-output GRAND (SO-GRAND) [31] employs a list decoder, which also keeps track of the probability of the correct ML codeword being in the accumulated list of candidate codewords at any given moment, allowing to dynamically adjust the list size by defining a threshold for that probability. Symbol-level GRAND has been recently proposed in [32] for single-input single-output (SISO) block fading channels, of which the additive Gaussian noise (AWGN) channel is a special case. Symbol-level GRAND attains significantly faster decoding than the original bit-level GRAND. In [2], the authors have suggested modifying GRAND to use knowledge about the adopted modulation scheme for channels with memory. Symbol-level GRAND takes a different approach: it relies on a closed-form expression for the probability that the input stream of bits contains a specific combination of bit strings representing various constellation symbols. These constellation symbols have different numbers of nearest and next-nearest neighbors. When the transmission is done over a block fading channel, the expression allows to order the error patterns according to their likelihood.

With the aim of attaining the URLLC objectives, this work proposes a transmission scheme that integrates RLC encoding and symbol-level GRAND, aided by antenna sorting, into a mMIMO spatial multiplexing system. Both zero-forcing (ZF) detection and minimum mean square error (MMSE) detection are considered, however, as it would be expected in mMIMO setups, MMSE provides no benefit regarding performance or complexity reduction in respect to the ZF detection filter. While RLCs cater for the sought-after high reliability and GRAND offers reduced decoding complexity, mMIMO techniques enable high spectral efficiency through spatial multiplexing. We explain that symbol-level GRAND can be directly extended to mMIMO, if strong channel hardening (CH) conditions are assumed. Furthermore, we show that the considered mMIMO system can cope with adverse CH conditions, if the symbols at the output of the linear detector are ordered according to their reliability, which can be derived from channel state information (CSI).

The optimized re-ordering of symbols can be seen as an antenna sorting problem. Antenna sorting has been known to greatly impact the detection performance of MIMO systems that use a small number of spatial streams. Optimal antenna sorting strategies that rely on the notion of the *effective* signal-to-noise ratio (SNR) of a stream at the output of the MIMO detector [33], [34] have been devised for different MIMO detection methods. For example, antenna sorting was used to increase the performance of V-BLAST detectors [34], or to simultaneously improve the performance and reduce the complexity of sphere decoders [35]. Any extra information regarding the a priori probability of the error patterns should be used, and that concept is at the core of the proposals in this paper. In this paper, we use the effective SNR after ZF or MMSE detection filters as a sorting metric akin to the reliability of the QAM symbols, which carry the bit strings that build up the codewords. The proposed antenna sorting method further reduces the complexity of symbol-level GRAND. Moreover, we introduce a constraint based on a carefully selected row of the parity-check matrix to discard incompatible error patterns. While related ideas appear in [36], [37], our method offers a simpler and more easily implementable alternative with minimal modification to the decoder, which, combined with antenna sorting, can reduce the complexity by more than 90% in some configurations when compared to bit-level GRAND.

In summary, the contributions of the paper are as follows:

1) Symbol-level GRAND, introduced in [32] for RLCs combined with Gray-coded square M-QAM for transmission over SISO block Rayleigh fading channels, has been abstracted into a generalized framework that can be applied to any M-ary modulation scheme for which a constellation can be defined.
2) Antenna sorting has been proposed as a simple technique that can be easily integrated into symbol-level GRAND to make it fit for MIMO channels. The sorting that results from MMSE detection is compared with the one arising from ZF detection, and it is shown that ZF is capable of offering equal performance and complexity reduction at any SNR in mMIMO scenarios.
3) A general bit-level lightweight membership test has been developed that can be applied to any GRAND decoder. A study of this technique is presented that leads to finding the rules that maximize the number of eliminated error patterns to be tested. This proposal is combined with symbol-level sorted GRAND to filter out a large number of considered error patterns that would have otherwise been subjected to the standard – more computationally intensive – membership test, thus reducing the overall computational requirements of the noise-guessing decoder.
4) The model of a system, which encompasses a transmitter employing RLC encoding, M-ary modulation and spatial multiplexing, a mMIMO channel and a receiver that uses ZF or MMSE detection and GRAND, has been created. It is shown that block error rate (BLER) measurements have been obtained for the case when the GRAND algorithm is either bit-level GRAND [16] or symbol-level GRAND [32], and is optionally combined with antenna sorting. The ideal case of perfect channel hardening (PCH) has been used as a lower-bound benchmark. The impact of the chosen GRAND algorithm on complexity has also been investigated.

The paper starts by describing the system model in detail in Section II. Section III shows how symbol reliability can be obtained from CSI and how antenna sorting should be implemented. Section IV generalizes symbol-level GRAND for any M-ary modulation scheme and integrates it with antenna sorting. Section V introduces a lightweight membership test that constrains the number of error patterns queried by symbol-level GRAND and, consequently, reduces computational cost. Section VI describes how symbol-level GRAND can be adapted to Gray-coded square M-QAM in the case of PCH, and discusses memory requirements. Section VII shows performance and complexity results of the proposed scheme, and Section VIII summarizes key conclusions.

II. System model for coded massive MIMO

A coded uplink massive MIMO system is considered, making use of an RLC encoder at the transmitter and symbol-level GRAND at the receiver. A block of k information bits is mapped onto a n-bit codeword and sent "over the air" via spatial multiplexing, as illustrated in Fig. 1. This process may be repeated when transmitting longer information streams by dividing the bit stream into blocks of size k.

A. RLC encoding and spatial multiplexing with mMIMO

A block $\mathbf{a}$ of k i.i.d. information bits is linearly encoded into a codeword $\mathbf{x}_\mathrm{b}$ of length n using a systematic binary RLC with rate $R = k/n$, denoted by RLC (n, k). The RLC (n, k) defines a codebook $\mathcal{C}$ with $2^k = 2^{nR}$ codewords of length n, which constitutes a linear subspace of the discrete vector space $\mathbb{F}_2^n$. Although the minimum Hamming distance between two codewords determines the error correction capability of (random or nonrandom) linear block codes (LBCs) at high-SNR, in very noisy conditions the minimum distance is not as relevant in determining the performance of *long* RLCs. In fact, when the objective is to approach the Shannon limit, the code needs to decode beyond the minimum distance [38, Ch.13]. This work focuses on the finite blocklenght, using short RLCs. The RLC (n, k) is described by a *random* binary generator matrix $\mathbf{M} \in \mathbb{F}_2^{k\times n}$, which acts as the basis matrix for the code subspace, such that $\mathcal{C} = \{\mathbf{x}_\mathrm{b} = \mathbf{aM} : \mathbf{a} \in \mathbb{F}_2^k\}$. The generator matrix of a systematic RLC is of the form $\mathbf{M} = [\,\mathbf{I}_k \mid \mathbf{P}\,]$, where $\mathbf{P} \in \mathbb{F}_2^{k\times(n-k)}$ is a random binary matrix, and $\mathbf{I}_k$ is the $k \times k$ identity matrix responsible for the systematic part of the encoding. The corresponding parity-check matrix will take the form $\mathbf{C} = \left[\,\mathbf{P}^T \mid \mathbf{I}_{n-k}\,\right] \in \mathbb{F}_2^{(n-k)\times n}$.

The n bits, $b_1, \ldots, b_n$, of a codeword are input to a M-ary constellation. The mapper divides the sequence of n bits into L strings of $\log_2(M)$ bits, that is, $L = n/\log_2(M)$, and maps the L strings onto L complex-valued symbols, $s_1, \ldots, s_L$, taken from the alphabet $\mathcal{A} \in \mathbb{C}$ of the M-ary constellation. The cardinality of $\mathcal{A}$ is $|\mathcal{A}| = M$. We denote the n-bit codeword and the sequence of L modulated symbols by $\mathbf{x}_\mathrm{b} = [b_1, \ldots, b_n]$ and $\mathbf{x}_\mathrm{s} = [s_1, \ldots, s_L]^T$, respectively. Furthermore, we denote by $\mathcal{S}(s_i)$ the string of $\log_2(M)$ bits that has been mapped onto symbol s_i. Thus, the codeword $\mathbf{x}_\mathrm{b}$ can also be written as $\mathbf{x}_\mathrm{b} = [\mathcal{S}(s_1), \ldots, \mathcal{S}(s_L)]$. If E_b represents the energy per information bit, then $(k/n)E_b$ is the energy per codeword bit, and $\log_2(M)(k/n)E_b$ is the energy per string of $\log_2(M)$ bits, which also corresponds to the energy per symbol, $\mathbb{E}\{|s_i|^2\}$.

The system can be designed to allow the transmission of N_c codewords in each MIMO channel used. This implies that, when a specific cardinality M is employed for the modulation, the number of transmit antennas is $N_T = N_c L$. Without loss of generality, and to keep the notation simple, we will describe the system for $N_c = 1$, where one MIMO burst transmitted from the N_T antennas contains one codeword only (i.e., $N_T = L$). Later, in subsection B, the generalization for $N_c > 1$ will be commented on. A system with $N_c < 1$ can also be made operational by adding buffers both at the transmitter and at the receiver, hence creating a full separation between the mMIMO physical layer and channel coding and decoding such that symbol-level GRAND only starts decoding when the L symbols corresponding to a codeword have been received.

The coded signal $\mathbf{x}_\mathrm{s}$ is transmitted over a MIMO Rayleigh fading channel, characterized by the matrix $\mathbf{H} \in \mathbb{C}^{N_T \times N_R}$, where $N_R \gg N_T$ is the number of antennas fitted at the receiver. The received signal $\mathbf{y}_\mathrm{s} = [y_1, ..., y_{N_R}]^T$ is given by

$$\mathbf{y}_\mathrm{s} = \mathbf{H}\mathbf{x}_\mathrm{s} + \mathbf{n}, \tag{1}$$

where $\mathbf{n} = [n_1, ..., n_{N_R}]^T$ is the additive noise. The entries in both $\mathbf{H}$ and $\mathbf{n}$ are i.i.d. random variables taken from complex normal distributions: the ones in $\mathbf{H}$ are taken from $\mathcal{CN}(0,1)$, and those in $\mathbf{n}$ are taken from $\mathcal{CN}(0, \sigma_\mathrm{n}^2)$. The noise power at each receive antenna is $\sigma_\mathrm{n}^2 = N_0$, where N_0 is the unilateral power spectral density of the noise. In this framework, the ergodic SNR per transmit antenna (i.e., per layer) at the receiver is

$$\mathrm{snr} \triangleq \frac{\mathbb{E}\{|s_i|^2\}}{\sigma_\mathrm{n}^2} = \log_2(M)\,(k/n)\,(E_b/N_0)\,. \tag{2}$$

The symbols in $\mathcal{A}$ are normalized to unit average energy, so that $\mathbb{E}\{|s_i|^2\} = 1$, and therefore $\mathrm{snr} = \frac{1}{\sigma_\mathrm{n}^2}$. The $N_T \times N_R$ matrix $\mathbf{H}$ remains constant during the transmission of $\mathbf{x}_\mathrm{s}$ but changes independently from channel use to channel use.

B. Linear detection and symbol-level GRAND

The ZF and MMSE linear detection schemes are considered [34], [39]. The first amounts to applying the Moore-Penrose pseudo-inverse

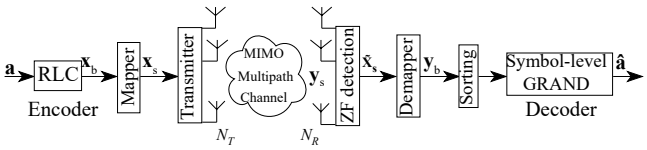

FIGURE 1. System model for coded mMIMO URLLC.

$$\mathbf{W}_\mathrm{ZF} = \left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{H}^H, \tag{3}$$

while the MMSE filter is given by

$$\mathbf{W}_\mathrm{MMSE} = \left(\mathbf{H}^H\mathbf{H} + \frac{1}{\mathrm{snr}}\mathbf{I}_{N_T}\right)^{-1}\mathbf{H}^H. \tag{4}$$

In the ZF case one obtains at the receiver

$$\mathbf{W}_\mathrm{ZF}\mathbf{y}_\mathrm{s} = \mathbf{I}_{N_T}\mathbf{x}_\mathrm{s} + \underbrace{\mathbf{W}_\mathrm{ZF}\mathbf{n}}_{\mathbf{u}}, \tag{5}$$

where $\mathbf{I}_{N_T}$ is the $N_T \times N_T$ identity matrix, and $\mathbf{u} \in \mathbb{C}^{N_T}$ denotes the new (now correlated) noise vector after ZF filtering. Although the performance of ZF detection is rather poor in symmetric MIMO, where $N_R = N_T$, it attains quasi-optimal performance in highly asymmetric MIMO, for example, when $N_R >> N_T$ in an uplink scenario [39]. In this scenario, which is considered in our system model, the instantaneous SNR for each channel realization approaches its ergodic value at each received data stream after ZF detection. At the same time, the large value of N_R boosts the receiver array gain.

After applying $\mathbf{W}_\mathrm{ZF}$ in (3) or $\mathbf{W}_\mathrm{MMSE}$ in (4), which can both be abstracted to $\mathbf{W}$, a quantization operation $\mathcal{Q}(\cdot)$ is made to the M-ary constellation to obtain the sequence of detected symbols $\tilde{\mathbf{x}}_\mathrm{s} = \mathcal{Q}(\mathbf{W}\mathbf{y}_s) = [\tilde{s}_1, \ldots, \tilde{s}_L]^T$, which is an estimate of $\mathbf{x}_\mathrm{s}$ corrupted by noise. The detected symbols $\tilde{s}_1, \ldots, \tilde{s}_L$ are demapped to bit strings $\mathcal{S}(\tilde{s}_1), \ldots, \mathcal{S}(\tilde{s}_L)$ and reconstruct a word of n bits, denoted by $\mathbf{y}_\mathrm{b}$. The relationship between the reconstructed word $\mathbf{y}_\mathrm{b}$ at the receiver and the codeword $\mathbf{x}_\mathrm{b}$ at the transmitter is $\mathbf{y}_\mathrm{b} = \mathbf{x}_\mathrm{b} \oplus \mathbf{e}_\mathrm{b}$, where $\mathbf{e}_\mathrm{b}$ is the error pattern that has corrupted the transmitted codeword. The operation $\oplus$ denotes modulo-2 addition.

The most common method of decoding LBCs, including RLCs, is syndrome decoding, which makes use of the parity-check matrix $\mathbf{C} \in \mathbb{F}_2^{(n-k)\times n}$ to generate the syndrome $\mathbf{s} = \mathbf{C}\mathbf{y}_\mathrm{b} \in \mathbb{F}_2^{(n-k)}$. Since all codewords verify $\mathbf{s} = \mathbf{0}$, this can be used as a simple *membership test*. Syndrome decoding achieves ML decoding, but a lookup table is required for the storage of possible syndromes and respective coset leaders. For example, suppose that we wish to correct received words that contain up to a threshold of w_th bit errors. The number of error patterns that need to be considered is given by $\sum_{t=0}^{w_\mathrm{th}} \binom{n}{t}$. However, the number of all possible syndromes is 2^{n-k}. For large values of w_th and high code rates, that is $R \to 1$ and thus $n \to k$, the relationship $\sum_{t=0}^{w_\mathrm{th}} \binom{n}{t} \gg 2^{(n-k)}$ holds, therefore the error correction capability of the code is limited because a wide variety of error patterns result in

the same syndrome. Choosing the coset leader associated with each particular syndrome depends on side information regarding the *a priori* probability of each error pattern. Over AWGN, the chosen coset leaders should be the error patterns with the lowest Hamming weight, which leads to ML decoding.

In the proposed transmission scheme, the word $\mathbf{y}_\mathrm{b}$ is input to symbol-level GRAND, which attempts to estimate $\mathbf{e}_\mathrm{b}$ and infer $\mathbf{x}_\mathrm{b}$ using $\hat{\mathbf{x}}_\mathrm{b} = \mathbf{y}_\mathrm{b} \oplus \hat{\mathbf{e}}_\mathrm{b}$, where $\hat{\mathbf{x}}_\mathrm{b}$ and $\hat{\mathbf{e}}_\mathrm{b}$ are estimates of $\mathbf{x}_\mathrm{b}$ and $\mathbf{e}_\mathrm{b}$, respectively. The first k of the n bits of the estimated codeword $\hat{\mathbf{x}}_\mathrm{b}$ form the block of decoded information bits $\hat{\mathbf{a}}$, as shown in Fig. 1.

III. Symbol reliability and antenna sorting

A. Effective post-processing SNR

After a ZF or MMSE filter, the decisions made by the quantizer $\mathcal{Q}(\cdot)$ to obtain $\hat{\mathbf{x}}_\mathrm{s}$ are perturbed by the modified noise vector $\mathbf{u}$ that appears in (5). One can show that the output SNR after ZF or MMSE detection of the N_T incoming signals streams depends on the instantaneous channel realization $\mathbf{H}$ in the following manner [33] [34, sec. 3.1.3 – 3.1.4]:

$$\begin{aligned} \mathrm{snr}_i^{(\mathrm{ZF})} &= \frac{\mathrm{snr}}{\left[(\mathbf{H}^H\mathbf{H})^{-1}\right]_{ii}} \\ &= \frac{1}{[\mathbf{G}^{-1}]_{ii}}\mathrm{snr} = g_i\ \mathrm{snr}, \quad 1 \le i \le N_T \end{aligned} \tag{6}$$

$$\mathrm{snr}_i^{(\mathrm{MMSE})} = \frac{\mathrm{snr}}{\left[(\mathbf{H}^H\mathbf{H} + \frac{1}{\mathrm{snr}}\mathbf{I}_{N_T})^{-1}\right]_{ii}} - 1, \quad 1 \le i \le N_T. \tag{7}$$

In the ZF case in (6), the g_i are defined as the inverses of the elements in the diagonal of $\mathbf{G}^{-1}$, for $i = 1, \ldots, N_T$. Note that $\mathbf{G} = \mathbf{H}^H\mathbf{H}$ is the Gram matrix of the lattice spanned by the columns of $\mathbf{H}$ (e.g., [40]). The value of each g_i in (6) should be as large as possible. In the case of a diagonal $\mathbf{G}$, that maximization happens for a $\mathbf{G}$ with large diagonal elements. If the energy spills over the diagonal, the elements in the diagonal get smaller due to energy conservation arguments. This corresponds to having the off-diagonal elements of $\mathbf{G}$ no longer close to zero due to non-orthogonality of the column vectors of $\mathbf{H}$. Note that a different definition of snr is used in [33], but that does not change the relation in (6).

Expressions (6) and (7) provide soft information about the reliability of each detected symbol, given the one-to-one relation between symbols and antenna streams. This information will be central to sorting the received symbols so that symbol-level GRAND can perform its guesswork of the transmitted symbols starting from the least reliable symbol to the most reliable one. While both expressions provide the absolute post-processing SNRs at each antenna, only the *relative* information about their magnitude ordering is required as a side information to be passed to the symbol-level GRAND decoder. It is known that in highly asymmetrical mMIMO configurations, MMSE delivers the same performance as ZF, while having a slightly more complex implementation. In our proposed receiver, it is important to determine whether the sorting provided by (6) and (7) has an impact on the symbol-level GRAND decoding complexity. At high snr, (6) tends to (7), and therefore the orders resulting from both should be equal. However, when snr decreases, the sorting orders could differ. The proportion of instances in which the ordered vectors produced by ZF matched those generated by MMSE was obtained and is shown in Fig. 2. The ordering vectors generated by sorting $\mathrm{snr}_i^{(\mathrm{ZF})}$ and $\mathrm{snr}_i^{(\mathrm{MMSE})}$ for the N_T antenna streams, by means of (6) and (7), were compared using 5000 channel realizations. The cumulative distribution function (CDF) shows that both resulting orders are the same at high SNR. As the SNR lowers, the ordered vectors start to differ. The cliff region of the CDF curves can be displaced to lower SNRs by applying an increasing number of receive antennas. The SNR range presented in Fig. 2, as well as the range of values for N_T and N_R, are covered in Section VII.

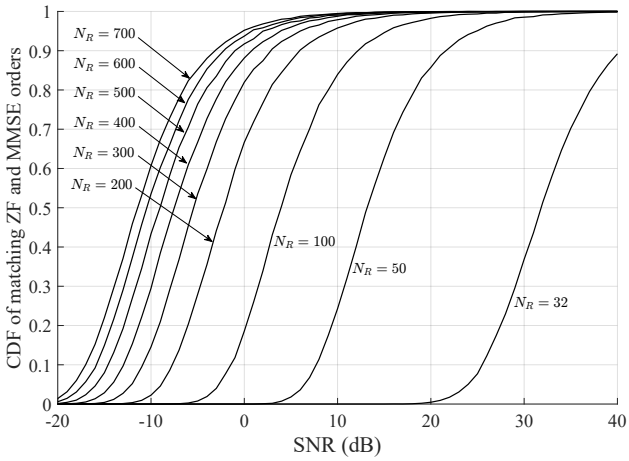

FIGURE 2. Cumulative distribution function of the number of exact matches between the ordered vectors defined by the sorting of the antennas when using ZF detection and when using MMSE detection filters. Different cases are represented for the configurations with $N_T = 32$ transmit antennas and a variable number of receive antennas, N_R. Simulation results obtained considering 5000 channel realizations at each SNR (computed at each integer value of the dB horizontal axis).

The sorting orders resulting from (6) and (7) are trivially the same as $\mathrm{snr} \to \infty$ because (7) tends to (6). To investigate how sorting is impacted for $\mathrm{snr} \to 0$, let us write the denominator in (7) as $\left(\mathbf{G} + \frac{1}{\mathrm{snr}}\mathbf{I}_{N_T}\right)^{-1} = \left(\frac{1}{\mathrm{snr}}\left(\mathbf{I}_{N_T} + \mathrm{snr}\,\mathbf{G}\right)\right)^{-1} = \mathrm{snr}\,(\mathbf{I}_{N_T} + \mathrm{snr}\,\mathbf{G})^{-1}$. Therefore, the first term in (7), the one responsible for the sorting order, is

$$\frac{\mathrm{snr}}{\left[(\mathbf{G} + \frac{1}{\mathrm{snr}}\mathbf{I}_{N_T})^{-1}\right]_{ii}} = \frac{1}{\left[(\mathbf{I}_{N_T} + \mathrm{snr}\,\mathbf{G})^{-1}\right]_{ii}}.$$

By using the Neumann series $(\mathbf{I}_{N_T} - \mathbf{G})^{-1} = \sum_{k=0}^{\infty}\mathbf{G}^k$, for low snr, one has $(\mathbf{I}_{N_T} + \mathrm{snr}\,\mathbf{G})^{-1} = \mathbf{I}_{N_T} - \mathrm{snr}\,\mathbf{G} + O(\mathrm{snr}^2)$. Consequently, when $\mathrm{snr} \to 0$, the first term in (7) tends to an identity matrix, turning equal all diagonal indexes i. Hence, at low SNR, the ZF sorting will be defined by the g_i gains, while the MMSE sorting becomes irrelevant. In the mMIMO

scenario, given that $\mathbf{G} \to \mathbf{I}_{N_T}$, the same effect will occur in ZF, eliminating the need for antenna sorting regardless of the linear detector used (see Section III.C). For a given snr, a larger N_R gives rise to a more orthogonal channel and therefore to a Gram matrix $\mathbf{G}$ that is closer to a scaled identity matrix. In that case, the denominators in (6) and (7) are more similar, leading to an increasing matching, as observed in the results in Fig. 2.

This section established that antenna sorting using ZF is equivalent to MMSE in mMIMO systems. For this reason, expression (6) will be used for the calculation of the SNR at the output of each receive antenna under the assumption of perfect CSI. A derivation of (6) is presented in Appendix A. The impact of imperfect CSI on the SNR expression is briefly explored and discussed in Appendix B.

B. Lattice geometry with a finite number of antennas

The geometry of ZF detection fully determines its detection performance. For ZF to approach ML detection using the Voronoi regions of the underlying real MIMO lattice, it is necessary that the so-called ZF detection region matches the Voronoi region with a low discrepancy (e.g., [34]). This is the fundamental cause for ZF detection becoming optimal as N_R increases. When $N_R \to \infty$ the lattice spanned by the columns of $\mathbf{H}$ would be a perfectly orthogonal lattice, and ZF would be optimal. Analytically, this effect can be captured by measuring the effect of the effective noise $\mathbf{u}$ in (5). The effect of the ZF filter on that noise power can be tracked by considering the autocorrelation matrix of the new noise $\mathbf{u} = \mathbf{W}_{\text{ZF}} + \mathbf{n}$, calculated as:

$$\begin{aligned}\mathbf{R}_{\mathbf{u}} &= \mathbb{E}\left\{\mathbf{u}\mathbf{u}^H\right\} = \mathbb{E}\left\{\left(\mathbf{W}_{\text{ZF}}\mathbf{n}\right)\left(\mathbf{W}_{\text{ZF}}\mathbf{n}\right)^H\right\} \\ &= \mathbb{E}\left\{\left(\mathbf{W}_{\text{ZF}}\mathbf{n}\right)\left(\mathbf{n}^H(\mathbf{W}_{\text{ZF}})^H\right)\right\} \\ &= \mathbf{W}_{\text{ZF}}\mathbb{E}\left\{\mathbf{n}\mathbf{n}^H\right\}(\mathbf{W}_{\text{ZF}})^H = \sigma_{\text{n}}^2\mathbf{W}_{\text{ZF}}(\mathbf{W}_{\text{ZF}})^H,\end{aligned} \tag{8}$$

where the autocorrelation of the original Gaussian noise, $\mathbb{E}\left\{\mathbf{n}\mathbf{n}^H\right\} = \mathbf{R}_{\mathbf{n}} = \sigma_{\text{n}}^2\mathbf{I}_{N_R}$, has been used. Replacing the Moore-Penrose pseudo-inverse from (3) in (8), and using the definition of the Gram matrix, it is possible to obtain

$$\mathbf{R}_{\mathbf{u}} = \sigma_{\text{n}}^2\left(\mathbf{H}^H\mathbf{H}\right)^{-1} = \sigma_{\text{n}}^2\mathbf{G}^{-1}. \tag{9}$$

From both (6) and (9), one can see that ZF detection always causes noise enhancement in the case of real-world channels (with a finite N_R). The noise amplification of ZF detection can be geometrically interpreted using lattices. Let $\mathcal{G} = \mathbb{Z} + i\mathbb{Z}$ denote the set of Gaussian integers. A complex lattice is defined as $\Lambda = \{\mathbf{H}\mathbf{z} : \mathbf{z} \in \mathcal{G}^{N_T\times 1}\}$. For a lattice basis $\mathbf{H} \in \mathbb{C}^{N_R\times N_T}$, the lattice has rank N_T, and lives in a N_R-dimensional space. The volume of the *fundamental region* of the lattice is $\text{vol}(\Lambda) = \sqrt{\det\left(\mathbf{H}^H\mathbf{H}\right)} = \sqrt{\det(\mathbf{G})}$. In the case of square matrices, this simplifies to $\text{vol}(\Lambda) = \det(\mathbf{H})$. In MIMO detection it is preferable to use the real-valued equivalent lattice, defined as $\Lambda_{\mathbb{R}} = \{\mathcal{H}\mathbf{z} : \mathbf{z} \in \mathbb{Z}^{2N_T\times 1}\}$, having rank $2N_T$, and living in $2N_R$ dimensions. It uses the equivalent *real-valued* basis $\mathcal{H} \in \mathbb{R}^{2N_R\times 2N_T}$, constructed from the complex basis $\mathbf{H}$ [34]. Noise amplification is reduced when a lattice's fundamental region is closer to orthogonal. To measure how orthogonal a lattice $\Lambda_{\mathbb{R}}$ is, one can use the so-called *orthogonality defect* (OD) [41], which for a lattice spanned by a real basis $\mathcal{H}$ is defined as:

$$od(\mathcal{H}) = \frac{\prod_{i=1}^{2N_T}\|\mathcal{H}(:,i)\|}{\text{vol}(\Lambda_{\mathbb{R}})}. \tag{10}$$

The value of $od(\mathcal{H})$ is always greater than or equal to one, and can only attain the unit if the columns of $\mathbf{H}$ are orthogonal to one another. We now use this metric to investigate how N_R and N_T influence the geometry of the mMIMO lattice and, therefore, how far from optimal ZF detection is. Fig. 3 shows how the OD evolves with N_R, for different values of N_T. The figure shows the domain of N_R of more practical significance and an overlaid graph depicts the OD asymptotic convergence to the unit value when the number of receive antennas tends to infinity. The OD is assessed by generating random samples of $\mathcal{H}$, with its real entries drawn from $\mathcal{N}(0, \frac{1}{2})$. This corresponds to generating $2N_T$ random Gaussian vectors in a vector space of N_R (real) dimensions, with the dimension of the vector space being much larger than the number of random vectors drawn (i.e., $N_R >> N_T$). When this happens, those vectors are mutually orthogonal with high probability. As expected, larger N_T necessitates having a larger N_R in order to maintain the same $od(\mathcal{H})$ value, and as N_R increases, the column vectors of $\mathcal{H}$ tend to be mutually orthogonal.

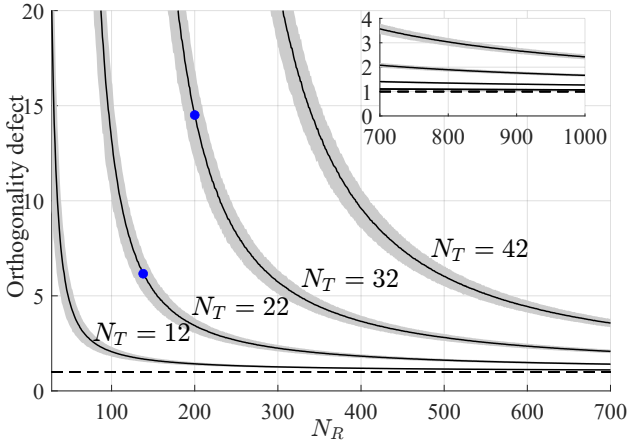

FIGURE 3. Evolution of $od(\mathcal{H})$ as a function of the number of receive antennas. The blue dots indicate the operating points of the two systems that will be assessed in Section VII that are closer to (but still far from) the PCH regime. The shaded region corresponds to two standard deviations of od.

C. Perfect channel hardening lower-bound

There is one specific (and ideal) circumstance in which noise amplification is prevented: when all the column vectors in $\mathbf{H}$ are mutually orthogonal. This occurs when N_T is fixed and $N_R \to \infty$, leading to the so-called *channel hardening effect* [42]. For a geometric interpretation of this property, one could consider N_T random Gaussian vectors living in a finite N_R-dimensional space. With high probability any pair

of the N_T vectors will be orthogonal to each other. With $N_R \to \infty$, this probability becomes 1. Let us consider a finite N_R and the special case of a channel matrix where an *ideal* MIMO channel is formed, i.e., a case where all columns of $\mathbf{H}$ are mutually orthogonal. In this case, the Gram matrix, which comprises all inner products $\mathbf{h}_i^H \mathbf{h}_j$, $i = 1, \dots N_R$, $j = 1, \dots N_T$, becomes a diagonal matrix of the form:

$$\mathbf{G} = \begin{bmatrix} \|\mathbf{h}_1\|^2 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \|\mathbf{h}_{N_T}\|^2 \end{bmatrix} = N_R \mathbf{I}_{N_T}, \tag{11}$$

given that $\|\mathbf{h}_j\|^2 = \sum_{i=1}^{N_R} |\mathbf{h}_{ij}|^2 = N_R$, for all the N_T vectors. By replacing (11) in (8) one gets that the autocorrelation of the noise after ZF, in the case of PCH, is

$$\mathbf{R}_{\mathbf{u}} = \frac{\sigma_{\mathrm{n}}^2}{N_R} \mathbf{I}_{N_T}, \tag{12}$$

and the power of $\mathbf{u}$ comes as

$$\|\mathbf{u}\|^2 = \mathrm{Tr}\left(\mathbf{R}_{\mathbf{u}}\right) = \frac{\sigma_{\mathrm{n}}^2 N_T}{N_R}. \tag{13}$$

It is now possible to establish the equivalent channel model if the $N_T \times N_R$ mMIMO configurations were to attain PCH at those (finite) dimensions:

$$\mathbf{W}_{\mathrm{ZF}} \mathbf{y}_{\mathrm{s}} = \mathbf{I}_{N_T} \mathbf{x}_{\mathrm{s}} + \mathbf{u}. \tag{14}$$

In this scenario, one has N_T independent parallel channels, where the effective noise becomes again a vector of independent Gaussian entries. Each of these N_T components of $\mathbf{u}$ has power $|u_i|^2 = \frac{\sigma_{\mathrm{n}}^2}{N_R}$, shedding light on the benefit of having a larger receiver array: with $N_R \to \infty$ there is a regression to the mean, and the effective noise power vanishes. Note that the snr in (14) is the *input* SNR, at each receive antenna before any baseband processing takes place. This asymptotic regime leads to a uniform *post-processing* $\mathrm{snr}_i^{(ZF)}$ across the N_T spatially multiplexed layers. In that limit, the reliability of all symbols is equal, and therefore sorting would bring no benefit.

IV. Generalized symbol-level GRAND with antenna sorting

Symbol-level GRAND was originally proposed for a SISO block Rayleigh fading channel [32], thus it cannot be directly applied to a mMIMO system. Taking in consideration the analysis made in the previous section, this section outlines the principles of symbol-level GRAND, generalizes them for any modulation scheme and describes how this method can be integrated into the mMIMO setup by incorporating soft information emanating from the ZF detector.

A. Sorting error patterns guided by the constellation structure

As described in Section II, the n-bit codeword $\mathbf{x}_{\mathrm{b}}$ can be expressed as a sequence of L strings, i.e., $\mathbf{x}_{\mathrm{b}} = [\mathcal{S}(s_i)]_{i=1}^{L}$, where s_i is a symbol of the M-ary modulation scheme and $\mathcal{S}(s_i)$ is the string of $\log_2(M)$ bits that has been mapped onto s_i. The demodulator outputs $\mathbf{y}_{\mathrm{b}} = [\mathcal{S}(\tilde{s}_i)]_{i=1}^{L}$, where the hard-detected symbols $\tilde{s}_1, \dots, \tilde{s}_L$ are demapped onto bit strings $\mathcal{S}(\tilde{s}_1), \dots, \mathcal{S}(\tilde{s}_L)$, respectively. The relationship between $\mathcal{S}(\tilde{s}_i)$ at the receiver and $\mathcal{S}(s_i)$ at the transmitter is $\mathcal{S}(\tilde{s}_i) = \mathcal{S}(s_i) \oplus e_i$, for $i = 1, \dots, L$, where e_i is the error string of length $\log_2(M)$ that altered the i-th bit string $\mathcal{S}(s_i)$. The received word $\mathbf{y}_{\mathrm{b}}$ assumes the form:

$$\begin{aligned} \mathbf{y}_{\mathrm{b}} &= [\mathcal{S}(\tilde{s}_i)]_{i=1}^{L} \\ &= [\mathcal{S}(s_i) \oplus e_i]_{i=1}^{L} \\ &= [\mathcal{S}(s_i)]_{i=1}^{L} \oplus [e_i]_{i=1}^{L} \\ &= \mathbf{x}_{\mathrm{b}} \oplus \mathbf{e}_{\mathrm{b}}, \end{aligned} \tag{15}$$

where $\mathbf{e}_{\mathrm{b}} = [e_i]_{i=1}^{L}$ is the error pattern that corrupted the transmitted codeword $\mathbf{x}_{\mathrm{b}}$. The objective of algorithms based on GRAND is to obtain an estimate of $\mathbf{e}_{\mathrm{b}}$, denoted by $\hat{\mathbf{e}}_{\mathrm{b}}$, and then add it to $\mathbf{y}_{\mathrm{b}}$ to estimate $\mathbf{x}_{\mathrm{b}}$, that is, $\hat{\mathbf{x}}_{\mathrm{b}} = \mathbf{y}_{\mathrm{b}} \oplus \hat{\mathbf{e}}_{\mathrm{b}}$ provided that $\hat{\mathbf{x}}_{\mathrm{b}} \in \mathcal{C}$.

Bit-level GRAND [16] keeps generating error patterns $\hat{\mathbf{e}}_{\mathrm{b}}$ in descending order of likelihood until an error pattern that satisfies the test $\mathbf{y}_{\mathrm{b}} \oplus \hat{\mathbf{e}}_{\mathrm{b}} \in \mathcal{C}$ is identified. The likelihood of an error pattern is assumed to be a monotonically decreasing function of its Hamming weight. In order to reduce complexity, bit-level GRAND introduced the notion of the abandonment threshold w_{th}. The generation of error patterns is abandoned when all error patterns of weight less than or equal to w_{th} have been tested and a valid codeword has not been found, i.e., $\mathbf{y}_{\mathrm{b}} \oplus \hat{\mathbf{e}}_{\mathrm{b}} \notin \mathcal{C}$ for every $\|\hat{\mathbf{e}}_{\mathrm{b}}\|_1 \leq w_{\mathrm{th}}$, where $\|\hat{\mathbf{e}}_{\mathrm{b}}\|_1$ denotes the Hamming weight of $\hat{\mathbf{e}}_{\mathrm{b}}$.

In symbol-level GRAND [32], the condition $\mathbf{y}_{\mathrm{b}} \oplus \hat{\mathbf{e}}_{\mathrm{b}} \in \mathcal{C}$ is replaced by the equivalent expression $[\mathcal{S}(\tilde{s}_i) \oplus \hat{e}_i]_{i=1}^{L} \in \mathcal{C}$, where $\hat{e}_i$ is the i-th error string of the estimated error pattern, i.e., $\hat{\mathbf{e}}_{\mathrm{b}} = [\hat{e}_i]_{i=1}^{L}$. Differently from bit-level GRAND, symbol-level GRAND does not generate and verify every realization of $\hat{\mathbf{e}}_{\mathrm{b}}$ for increasing Hamming weight; instead, it creates and tests only realizations of $\hat{\mathbf{e}}_{\mathrm{b}}$, which are composed of error strings that are more likely to occur, as dictated by the properties of the constellation diagram of the modulation scheme and the method for mapping bit strings onto symbols.

The concept of symbol-level GRAND was introduced in [32] through the lens of Gray-coded square M-QAM, as the analysis and expressions therein were tailored to that particular modulation scheme. In this section, we derive a generalized expression for the probability of occurrence of an error pattern when symbol-level GRAND is applied to any M-ary modulation scheme. As in [32], we will initially study the case where the L received symbols have been affected by the same fading coefficient and, therefore, have the same reliability. We will then present a simple add-on to symbol-level GRAND to enable the decoding of symbols that have different reliabilities, as in the considered mMIMO setup.

Let $[L_1 \;\; L_2 \dots L_\vartheta]$ describe the structure of an error pattern, where L_w denotes the number of error strings in the

TABLE 1. Examples of error patterns with structure $[L_1\ L_2]$ for $M = 16$ and $n = 16$.

Structure $[L_1\ L_2]$	Examples of error patterns having structure $[L_1\ L_2]$	Weight $(L_1 + 2L_2)$
[2 0]	Example 1: 0010 – 1000 – 0000 – 0000 Example 2: 0000 – 0001 – 0001 – 0000 Example 3: 0100 – 0000 – 0000 – 0010	2
[0 1]	Example 1: 0000 – 0011 – 0000 – 0000 Example 2: 0000 – 0000 – 1001 – 0000 Example 3: 1100 – 0000 – 0000 – 0000	2
[1 1]	Example 1: 0101 – 0001 – 0000 – 0000 Example 2: 0100 – 0000 – 0000 – 1100 Example 3: 0000 – 0000 – 0110 – 1000	3
[2 1]	Example 1: 0000 – 1000 – 0011 – 0100 Example 2: 0110 – 0001 – 0001 – 0000 Example 3: 0010 – 0000 – 0100 – 1010	4

error pattern that have Hamming weight w, for $w = 1, \ldots, \vartheta$, $1 \leq \vartheta \leq \log_2(M)$ and $\sum_{w=1}^{\vartheta} L_w \leq L$. The Hamming weight of the error pattern is given by $\sum_{w=1}^{\vartheta} wL_w$. Table 1 shows examples of error patterns of length $n = 16$ composed of $L = 4$ error strings, each having length $\log_2(M) = 4$. For clarity, error strings in an error pattern have been separated by dashes. Error strings of weight up to $\vartheta = 2$ are considered in Table 1, thus the structure of an error pattern can be summarized by $[L_1\ L_2]$. For example, the three error patterns listed at the bottom of Table 1 have structure $[L_1\ L_2] = [2\ 1]$, i.e., each error pattern consists of $L_1 = 2$ error strings of weight 1 (displayed using blue typeface) and $L_2 = 1$ error string of weight 2 (rendered in red lettering). In all three cases, the weight of the error pattern is $L_1 + 2L_2 = 4$.

If the fading coefficient of the channel is constant over a block of L symbols, error patterns that have the same structure $[L_1\ L_2 \ldots L_\vartheta]$, have the same probability of occurrence denoted by $P(L_1, L_2, \ldots, L_\vartheta)$. Symbol-level GRAND relies on an expression for $P(L_1, L_2, \ldots, L_\vartheta)$ to calculate the likelihood of every possible structure – or every structure that satisfies $\sum_{w=1}^{\vartheta} wL_w \leq w_{\text{th}}$ if we wish to bound from above the Hamming weight of the considered error patterns. Error patterns are then generated and tested for each structure, from the most likely structure to the least likely structure.

The error strings $\hat{e}_1, \ldots, \hat{e}_L$ that compose an error pattern do not take values uniformly at random, but depend on the transmitted symbols $s_1, \ldots, s_L$ and the noise power that led to the detection of symbols $\tilde{s}_1, \ldots, \tilde{s}_L$ at the receiver, given that $\hat{e}_i = \mathcal{S}(s_i) \oplus \mathcal{S}(\tilde{s}_i)$ for $i = 1, \ldots, L$. To facilitate the derivation of an expression for $P(L_1, L_2, \ldots, L_\vartheta)$, we:

- Define B *neighborhoods* for each symbol. Members of a neighborhood of a symbol are other symbols of the constellation diagram that are adjacent to that symbol and have the same Euclidean distance from it.
- Define T *types*, whereby symbols of the same type share the same neighborhood characteristics in terms of size, Euclidean distance, and also Hamming distance.

Fig. 4 shows examples of neighborhoods and symbol types for the 16-ary regular triangular QAM (TQAM) [43], [44], also referred to as regular hexagonal QAM [45], and for Gray-coded square 16-QAM. In the case of 16-ary regular TQAM, the nearest neighbors of a symbol are equidistant from that symbol and surround it, thus forming a single neighborhood for that symbol ($B = 1$). Notice that if s_i is transmitted and the detected symbol $\tilde{s}_i$ is in the neighborhood of s_i, the Hamming distance between $\mathcal{S}(s_i)$ and $\mathcal{S}(\tilde{s}_i)$ corresponds to the Hamming weight of the error string $\hat{e}_i$. Depending on the size of a neighborhood and the Hamming distance between a symbol and each of its neighbors, the 16 symbols can be classified into $T = 6$ types. If each type is arbitrarily assigned a unique number $j \in \{1, \ldots, T\}$, then t_j denotes the number of symbols of type j, where $\sum_{j=1}^{T} t_j = M$. For 16-ary regular TQAM, we have $t_1 = 2$, $t_2 = 2$, $t_3 = 4$, $t_4 = 2$, $t_5 = 2$ and $t_6 = 4$. In the case of Gray-coded square 16-QAM, the symbols that surround each symbol can be divided into $B = 2$ neighborhoods based on their Euclidean distance from that symbol. Each symbol can then be categorized into one of $T = 3$ types, where $t_1 = 4$, $t_2 = 8$ and $t_3 = 4$, as illustrated in Fig. 4(b).

Using the concepts of neighborhoods and types, the probability that an error pattern or, equivalently, a sequence of L error strings has structure $[L_1\ L_2 \ldots L_\vartheta]$ can be expressed as a function of the probability that L symbols classed into T types are selected uniformly at random for transmission, L' of them are received incorrectly, where $L' = \sum_{w=1}^{\vartheta} L_w$, and the remaining $L - L'$ symbols are received without errors. More specifically, we can write:

$$P(L_1, L_2, \ldots, L_\vartheta) = \\ = \frac{1}{M^L} \sum_{\sum_{j=1}^{T} \ell_j = L} \left\{ \binom{L}{\ell_1, \ldots, \ell_T} \prod_{j=1}^{T} t_j^{\ell_j} \times \right. \\ \left. \times \sum_{\substack{\sum_{j=1}^{T} \ell'_{j,0} = L - L' \\ \sum_{j=1}^{T} \ell'_{j,w} = L_w \\ \forall w \in \{1, \ldots, \vartheta - 1\}}} \left[\prod_{j=1}^{T} \binom{\ell_j}{\ell'_{j,0}, \ldots, \ell'_{j,\vartheta}} \prod_{w=0}^{\vartheta} p_{j,w}^{\ell'_{j,w}} \right] \right\}. \quad (16)$$

The second line of (16) determines the fraction of sequences of L symbols, from a total of M^L possible sequences, that contain ℓ_j symbols of type j, for $j = 1, \ldots, T$. Each of the ℓ_j positions in the sequence can be occupied by any of the t_j symbols of that type, resulting in $t_j^{\ell_j}$ possible outcomes. The third line of (16) enumerates all possible ways that $\ell'_{j,w}$ error strings of weight w can change any $\ell'_{j,w}$ of the ℓ_j transmitted symbols of type j into a neighboring symbol with probability $p_{j,w}^{\ell'_{j,w}}$, for $j = 1, \ldots, T$ and $w = 0, \ldots, \vartheta$. An error string of weight 0 leaves a transmitted symbol unchanged, thus the first condition $\sum_{w=1}^{T} \ell'_{j,0} = L - L'$

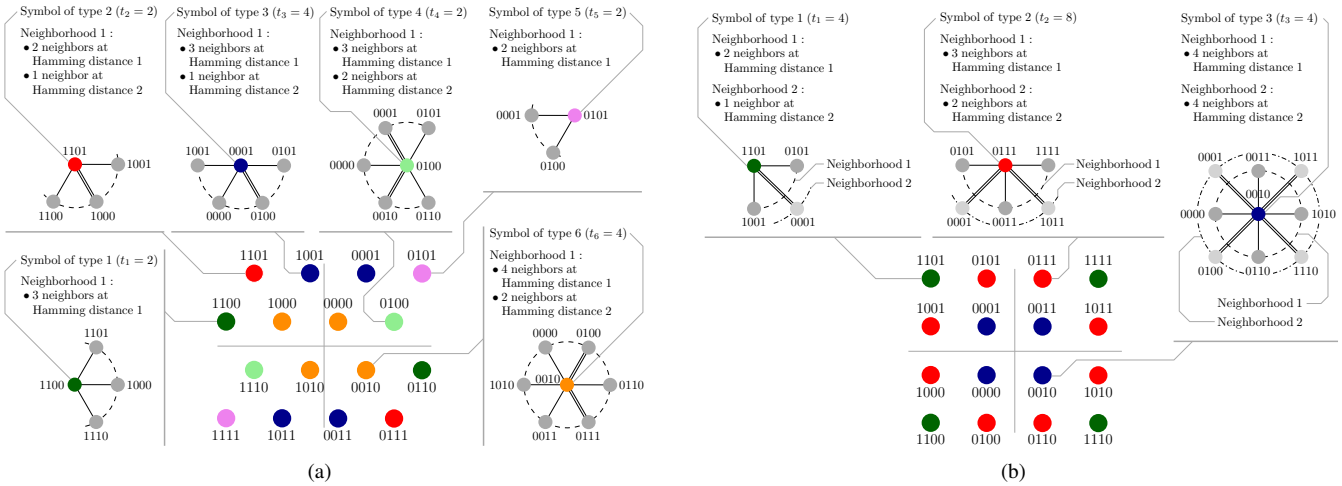

FIGURE 4. Constellation diagrams of (a) 16-ary regular TQAM and (b) Gray-coded square 16-QAM. In each case, every symbol has been classified into a type based on the number of adjacent symbols, their Euclidean distance and their Hamming distance from that symbol. Symbols located at the same Euclidean distance from a symbol compose a neighborhood for that symbol. Symbols connected by a single line are separated by a Hamming distance of 1. The Hamming distance between symbols linked by a double line is 2. If a symbol is transmitted and a neighboring symbol is received, the Hamming weight of the error string that altered the binary value of the transmitted symbol is equal to the Hamming distance of the two symbols.

ensures that the number of weight-0 error strings matches the number of correctly received symbols. The second condition requires that all error strings of weight w add up to L_w for $0 < w \leq \vartheta - 1$. Given the values of $\ell'_{j,0}, \ldots, \ell'_{j,\vartheta-1}$, the value of $\ell'_{j,\vartheta}$ can be derived from $\ell'_{j,\vartheta} = \ell_j - \ell'_{j,0} - \ldots - \ell'_{j,\vartheta-1}$. Expressions for $p_{j,w}$ depend on the M-ary modulation scheme. Section VI applies (16) to the case of Gray-coded square M-QAM and provides expressions for $p_{j,w}$.

Derivation of (16) assumed that received symbols have the same reliability, which is the case when symbols are transmitted over a SISO block Rayleigh fading channel. The following section explains how CSI could guide symbol-level GRAND in prioritizing error patterns of the same structure when received symbols have different reliabilities, as in the considered mMIMO scenario.

B. Sorting error patterns guided by CSI

In a non-ideal mMIMO scenario with $od(\mathcal{H}) > 1$, the reliability of spatial streams could vary significantly among them. At the receiver, a first processing block should implement the antenna sorting, as previously discussed in Section III.

Without loss of generality, we will discuss the case with $N_c = 1$. This can be accomplished by inserting a permutation matrix $\mathbf{\Pi}$, which is a binary matrix whose columns are all columns of the identity $\mathbf{I}$ but placed in a different order. As it is well known, a permutation matrix is always an orthogonal matrix, and its inverse is its transpose: $\mathbf{\Pi}^{-1} = \mathbf{\Pi}^T$. These two matrices can be added before and after symbol-level GRAND, as presented in Fig. 5. The $g_1, \ldots, g_{N_T}$ gains in (6) are sorted in ascending order of magnitude, and the corresponding permutation matrix $\mathbf{\Pi}$ is created. The permuted symbols $\tilde{\mathbf{x}}_s^{(\Pi)} = \tilde{\mathbf{x}}_s \mathbf{\Pi}$ are fed to symbol-level GRAND, which will now not assume that error patterns of the same structure are equiprobable and test them

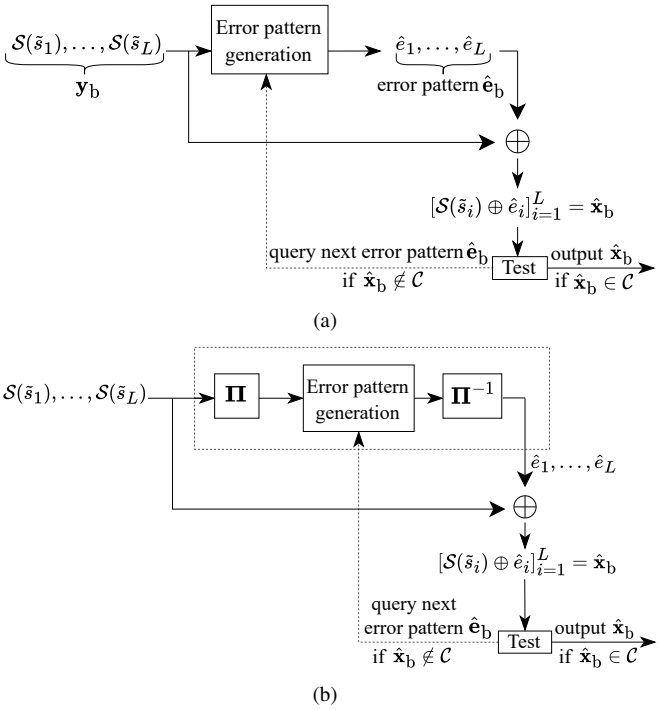

FIGURE 5. Types of symbol-level GRAND: (a) symbol-level GRAND proposed in [32], and (b) symbol-level GRAND with antenna sorting.

in an arbitrary order, as was the case in the previous section, but will test them in decreasing order of likelihood.

In the general case with $N_c > 1$ codewords per MIMO transmission, the set of the g_i, for $i = 1, \cdots, N_T = N_c L$, is partitioned in N_c subsets and an independent sorting process is applied to each one of those subsets. Note that, for a faster overall decoding time, these sorting processes can be implemented in parallel. Afterwards, each subset of L sorted

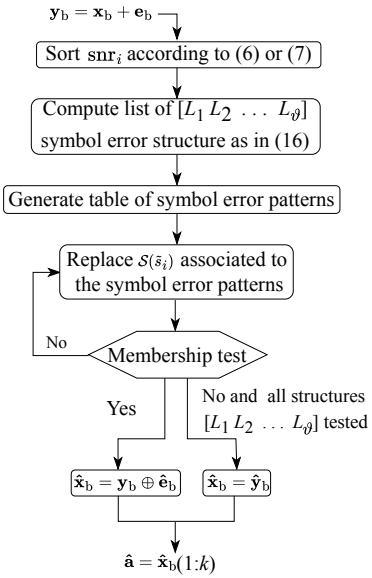

FIGURE 6. Symbol-level GRAND with antenna sorting after the demodulation and detection stages. (The final output of the algorithm makes use of MATLAB notation.)

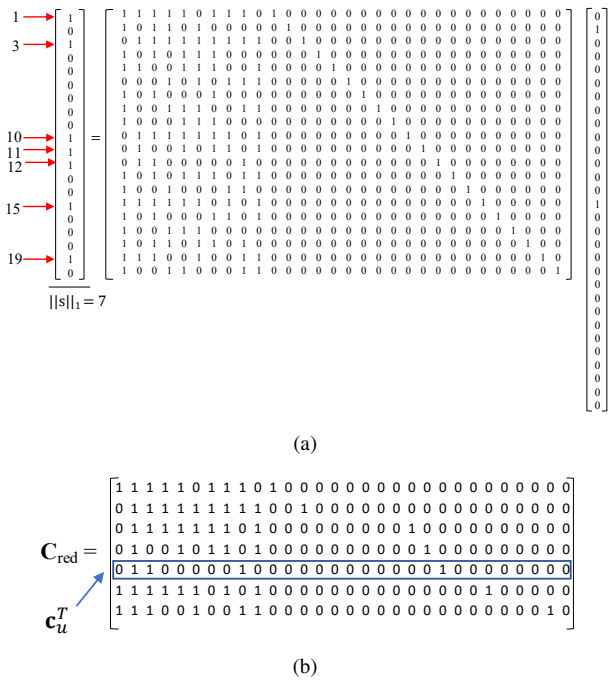

FIGURE 7. Example of the application of the lightweight membership test. (a) Full parity-check matrix $\mathbf{C}$ with the syndrome's flagged positions marked by red arrows, associated to the flagged rows of $\mathbf{C}$. (b) Reduced parity-check matrix $\mathbf{C}_\mathrm{red}$, where row $u = 5$ is the most unbalanced row, with $|\mathtt{S}_1^{(5)}| = 4$ and $|\mathtt{S}_0^{(5)}| = 26$.

symbols is passed on to the symbol-level GRAND, which independently decode each one of these N_c codewords. Likewise the sorting procedures, the decoding of each codeword can be performed in parallel, if further reduction of decoding latency is paramount. This may be done at the cost of having multiple symbol-level GRAND processors.

The flowchart of the proposed symbol-level GRAND with antenna sorting is presented in Fig. 6 (for the $N_c = 1$ case). The membership test creates a computational bottleneck in the estimation process that can be remedied by the introduction of a preliminary lightweight test, as explained in the following section.

V. Lightweight membership test

Upon reception of $\mathbf{y}_\mathrm{b}$, symbol-level GRAND with antenna sorting generates error patterns in descending order of likelihood and, in theory, selects the first error pattern $\mathbf{e}_\mathrm{b}$ that satisfies $\mathbf{y}_\mathrm{b} \oplus \hat{\mathbf{e}}_\mathrm{b} \in \mathcal{C}$. In practice, $\mathbf{y}_\mathrm{b} \oplus \hat{\mathbf{e}}_\mathrm{b}$ is not compared against every codeword in $\mathcal{C}$. Instead, the decoder calculates the syndrome vector $\mathbf{s} = \mathbf{C}\mathbf{y}_\mathrm{b}$ using $\mathbf{C}$, i.e., the $(n-k) \times n$ parity-check matrix of the (n, k) RLC. The algorithm in Fig. 6 keeps generating error patterns until an error pattern $\hat{\mathbf{e}}_\mathrm{b}$ that meets $\mathbf{C}\hat{\mathbf{e}}_\mathrm{b} = \mathbf{s}$ is found, which implies that $\mathbf{y}_\mathrm{b} \oplus \hat{\mathbf{e}}_\mathrm{b}$ is a member of $\mathcal{C}$. In this section, we propose a lightweight test, which requires the multiplication of a considered error pattern with just one of the $n-k$ rows of $\mathbf{C}$. Only if this test is passed, the product $\mathbf{C}\hat{\mathbf{e}}_\mathrm{b}$ is calculated, thus constraining the number of error patterns that are multiplied with every row of $\mathbf{C}$.

The proposed refinement takes advantage of the positions of the non-zero elements in the syndrome. The positions in $\mathbf{s}$ holding ones are referred to as *flagged positions*, which correspond to the support of $\mathbf{s}$, denoted by $\mathrm{supp}\,(\mathbf{s})$, and their count is the Hamming weight $\|\mathbf{s}\|_1$. The rows of $\mathbf{C}$ corresponding to these flagged positions are extracted to form a reduced matrix $\mathbf{C}_\mathrm{red}$. We use MATLAB notation to denote the i-th row of $\mathbf{C}_\mathrm{red}$ as $\mathbf{C}_\mathrm{red}(i,:)$, for $1 \le i \le ||\mathbf{s}||_1$, and we define the *set of ones*, as the set of positions that contain ones is the support of that row:

$$\mathtt{S}_1^{(i)} = \mathrm{supp}\,(\mathbf{C}_\mathrm{red}(i,:))\,. \tag{17}$$

The weight of each row of $\mathbf{C}_\mathrm{red}$ is $||\mathbf{C}_\mathrm{red}(i,:)||_1 = |\mathtt{S}_1^{(i)}|$. The remaining positions in $\mathbf{C}_\mathrm{red}(i,:)$, i.e., its positions containing zeros, form the *set of zeros* of that row, denoted as $\mathtt{S}_0^{(i)}$. Note that the union $\mathtt{S}_1^{(i)} \cup \mathtt{S}_0^{(i)}$ corresponds to all n positions.

Fig. 7 depicts an example, whereby an error pattern $\hat{\mathbf{e}}_\mathrm{b}$ produces the pre-computed syndrome $\mathbf{s}$ upon multiplication with the parity-check matrix $\mathbf{C}$, i.e., the error pattern $\hat{\mathbf{e}}_\mathrm{b}$ satisfies the membership test $\mathbf{s} = \mathbf{C}\hat{\mathbf{e}}_\mathrm{b}$ in Fig. 6 and will be selected for the estimation of the transmitted codeword. The red arrows in Fig. 7(a) identify the $||\mathbf{s}||_1$ flagged positions in $\mathbf{s}$, which point to $||\mathbf{s}||_1$ rows in the $(n-k) \times n$ parity-check matrix $\mathbf{C}$. These $||\mathbf{s}||_1$ rows in $\mathbf{C}$ can be stacked to form the $||\mathbf{s}||_1 \times n$ matrix $\mathbf{C}_\mathrm{red}$, as shown in Fig. 7(b). In this example, $k = 10$, $n = 30$ and $||\mathbf{s}||_1 = 7$.

Let $\mathbf{c}_u^T = \mathbf{C}_\mathrm{red}(u,:)$ denote an arbitrarily selected row of $\mathbf{C}_\mathrm{red}$. Based on the definition of $\mathbf{C}_\mathrm{red}$, if a candidate error pattern $\hat{\mathbf{e}}_\mathrm{b}$ satisfies the equality $\mathbf{s} = \mathbf{C}\hat{\mathbf{e}}_\mathrm{b}$, the product $\mathbf{C}_\mathrm{red}\,\hat{\mathbf{e}}_\mathrm{b}$ will give the $||\mathbf{s}||_1 \times 1$ all-ones vector, hence the equality $\mathbf{c}_u^T\hat{\mathbf{e}}_\mathrm{b} = 1$ will also be satisfied. Conversely, if $\hat{\mathbf{e}}_\mathrm{b}$ does not satisfy $\mathbf{c}_u^T\hat{\mathbf{e}}_\mathrm{b} = 1$ and, instead, results in $\mathbf{c}_u^T\hat{\mathbf{e}}_\mathrm{b} = 0$, then it

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3630551

will fail the membership test $\mathbf{s} = \mathbf{C}\hat{\mathbf{e}}_\mathrm{b}$. Therefore, the constraint $\mathbf{c}_u^T\hat{\mathbf{e}}_\mathrm{b} = 1$ can be used as a lightweight test; only error patterns that meet the constraint and pass the lightweight test will be subjected to the membership test, whereas all other error patterns will be filtered out. In the latter case, the proposed constraint reduces the complexity associated with the computation of $\mathbf{C}\hat{\mathbf{e}}_\mathrm{b}$ from $\mathcal{O}((n-k)n) = \mathcal{O}(n^2)$ to $\mathcal{O}(n)$, as multiplication is limited to a single row of $\mathbf{C}$.

The search space of all error patterns of weight $||\hat{\mathbf{e}}_\mathrm{b}||_1$ that will be considered for the membership test has size $\binom{n}{||\hat{\mathbf{e}}_\mathrm{b}||_1}$. Let α_0 be the number of error patterns that result in $\mathbf{c}_u^T\hat{\mathbf{e}}_\mathrm{b} = 0$ and will thus be eliminated by the lightweight test. On the other hand, let α_1 be the number of error patterns that meet the condition $\mathbf{c}_u^T\hat{\mathbf{e}}_\mathrm{b} = 1$ and, therefore, will pass the lightweight test and undergo the membership test. The relationship between α_0, α_1 and the size of the unconstrained search space of all vectors of weight $||\hat{\mathbf{e}}_\mathrm{b}||_1$ is

$$\alpha_0 + \alpha_1 = \binom{n}{||\hat{\mathbf{e}}_\mathrm{b}||_1}. \tag{18}$$

To derive an expression for α_0 and α_1, the conditions for $\mathbf{c}_u^T\hat{\mathbf{e}}_\mathrm{b}$ being 0 or 1, respectively, need to be first determined.

If we shift our attention back to Fig. 7(b), we will notice that the product of the selected row $\mathbf{c}_u^T$ and the error pattern $\hat{\mathbf{e}}_\mathrm{b}$ is 1 because only one of the two positions in $\hat{\mathbf{e}}_\mathrm{b}$ that hold non-zero elements is in $\mathtt{S}_1^{(u)} = \{2, 3, 9, 22\}$. If the two non-zero elements of $\hat{\mathbf{e}}_\mathrm{b}$ occupied positions that were both members of $\mathtt{S}_1^{(u)}$, or they both occupied any of the remaining $n - |\mathtt{S}_1^{(u)}| = 26$ positions, the product would have been 0 and the lightweight test would have failed.

In general, an error pattern of weight $||\hat{\mathbf{e}}_\mathrm{b}||_1$ will fail the lightweight test and will be eliminated from the search space of the membership test, if an *even* number of non-zero elements in the error pattern occupy any of the $|\mathtt{S}_1^{(u)}|$ positions listed in $\mathtt{S}_1^{(u)}$, while the rest of the non-zero elements occupy the remaining $n - |\mathtt{S}_1^{(u)}| = |\mathtt{S}_0^{(u)}|$ positions. Hence, the total number of error patterns that will be eliminated by the lightweight test is given by

$$\alpha_0 = \sum_{\rho=\rho_\mathrm{lo}}^{\rho_\mathrm{up}} \binom{|\mathtt{S}_1^{(u)}|}{2\rho} \binom{n - |\mathtt{S}_1^{(u)}|}{||\hat{\mathbf{e}}_\mathrm{b}||_1 - 2\rho}. \tag{19}$$

Expressions for the lower and upper bounds of the summation, i.e., ρ_lo and ρ_up, can be deduced from the conditions $0 \le 2\rho \le |\mathtt{S}_1^{(u)}|$ and $0 \le \left(||\hat{\mathbf{e}}_\mathrm{b}||_1 - 2\rho\right) \le n - |\mathtt{S}_1^{(u)}|$ that the two binomial coefficients in (19) are subject to. The upper bound is defined by having no more ones to distribute among the positions in $\mathtt{S}_1^{(u)}$ or because there are no more empty positions in $\mathtt{S}_1^{(u)}$ to fill in with ones. Conversely, the lower bound is related with the situation when very few ones are placed in $\mathtt{S}_1^{(u)}$ and the remaining ones exceed the number of existing positions in the $\mathtt{S}_0^{(u)}$ set. The interval where the two ranges overlap determines the domain $[\rho_\mathrm{lo}, \rho_\mathrm{up}]$ of ρ.

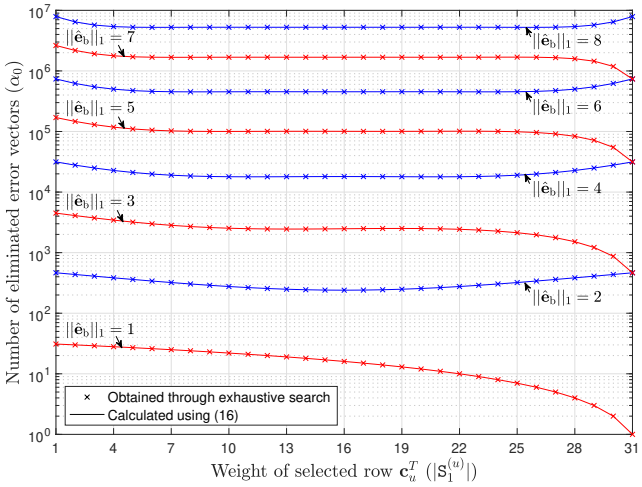

FIGURE 8. Number of error patterns that are eliminated by the lightweight test, as a function of the weight of the selected row $\mathbf{c}_u^T$. Error patterns of length $n = 32$ and weight $1 \le ||\hat{\mathbf{e}}_\mathbf{b}||_1 \le 8$ have been considered. Markers (×) correspond to values obtained through an exhaustive search, whereas solid lines (—) correspond to values obtained from (19).

We obtain

$$\rho_\mathrm{lo} = \max\left\{0, \left\lceil \frac{|\mathtt{S}_1^{(u)}| + ||\hat{\mathbf{e}}_\mathrm{b}||_1 - n}{2} \right\rceil\right\} \text{ and}$$
$$\rho_\mathrm{up} = \min\left\{\left\lfloor \frac{|\mathtt{S}_1^{(u)}|}{2} \right\rfloor, \left\lfloor \frac{||\hat{\mathbf{e}}_\mathrm{b}||_1}{2} \right\rfloor\right\},$$

where $\lceil t \rceil$ and $\lfloor t \rfloor$ are operators that round t up or down, respectively, to the nearest integer.

Expression (19) has been plotted in Fig. 8 in terms of $|\mathtt{S}_1^{(u)}|$, for $1 \le ||\hat{\mathbf{e}}_\mathrm{b}||_1 \le 8$ and $n = 32$. Similar trends emerge for different values of n. An exhaustive search has also been carried out to count all error patterns that do not satisfy $\mathbf{c}_u^T\hat{\mathbf{e}}_\mathrm{b} = 1$, and compare the total with the value of α_0 obtained from (19) to visually confirm the validity of (19) in Fig. 8. We can infer from the plotted curves that selecting a row $\mathbf{c}_u^T$ out of the $||\mathbf{s}||_1$ rows of $\mathbf{C}_\mathrm{red}$ – not arbitrarily but based on its weight – could maximize the value of α_0. As shown in Fig. 8, for *even* values of $||\hat{\mathbf{e}}_\mathrm{b}||_1$, choosing the row of $\mathbf{C}_\mathrm{red}$ with either the lowest or the highest weight, i.e., with the greatest imbalance in the number of zero and non-zero elements, will maximize the number of error patterns that the lightweight test eliminates. The index u of this row can be determined from

$$\begin{aligned} u &= \underset{1 \le i \le ||\mathbf{s}||_1}{\arg\min} \left\{|\mathtt{S}_1^{(i)}|, |\mathtt{S}_0^{(i)}|\right\} \\ &= \underset{1 \le i \le ||\mathbf{s}||_1}{\arg\min} \left\{|\mathtt{S}_1^{(i)}|, n - |\mathtt{S}_1^{(i)}|\right\}, \quad \text{for } ||\hat{\mathbf{e}}_\mathrm{b}||_1 \text{ even.} \end{aligned} \tag{20}$$

If multiple rows have the same weight, a value for u is selected at random from the tied candidates. On the other hand, for *odd* values of $||\hat{\mathbf{e}}_\mathrm{b}||_1$, choosing the row of $\mathbf{C}_\mathrm{red}$ with the lowest weight has the potential to maximize the

value of α_0. In this case, the index u can be obtained from

$$u = \underset{1 \le i \le ||\mathbf{s}||_1}{\arg\min} \left\{ |\mathrm{S}_1^{(i)}| \right\}, \quad \text{for } ||\hat{\mathbf{e}}_\mathrm{b}||_1 \text{ odd.} \tag{21}$$

The size of the constrained and, thus, reduced search space of error patterns of weight $||\hat{\mathbf{e}}_\mathrm{b}||_1$ for the *full* membership test can be derived from (18), if we solve for α_1 and substitute α_0 with the sum in (19). Note that in the case of systematic codes, due to the existence of $n-k-1$ zeros in the $\mathbf{I}_{n-k}$ submatrix in $\mathbf{C}$, the most unbalanced row selected by (20) often ends up corresponding to the row with the fewest ones, selected by (21).

VI. Analysis in the perfect channel hardening limit

The system model considered in this paper and presented in Section II focuses on Gray-coded square M-QAM. Symbol-level GRAND with antenna sorting tests error patterns of structure $[L_1\ L_2]$ and weight $L_1+2L_2 \le w_\mathrm{th}$. An expression for the likelihood of a structure $[L_1\ L_2]$, which will be used to order all structures prior to the generation and testing of error patterns, can be obtained from (16) for $\vartheta = 2$, $T = 3$, $t_1 = 4$, $t_2 = 4(\sqrt{M}-2)$, $t_3 = (\sqrt{M}-2)^2$ and $L' = L_1+L_2$, as follows:

$$\begin{aligned} P(L_1, L_2) \approx & \\ \frac{1}{M^L} \sum_{\ell_1+\ell_2+\ell_3=L} & \left\{ \binom{L}{\ell_1, \ell_2, \ell_3} 4^{\ell_1+\ell_2} \left(\sqrt{M}-2\right)^{\ell_2+2\ell_3} \right. \\ \times \sum_{\substack{\ell'_{1,0}+\ell'_{2,0}+\ell'_{3,0} = L-L' \\ \ell'_{1,1}+\ell'_{2,1}+\ell'_{3,1} = \ \ L_1}} & \left[\prod_{j=1}^{3} \binom{\ell_j}{\ell'_{j,0}, \ell'_{j,1}, \ell'_{j,2}} \times \right. \\ \left. \left. \times \prod_{w=0}^{2} p_{j,w}^{\ell'_{j,w}} \right] \right\}, & \end{aligned} \tag{22}$$

which is in agreement with the expression derived in [32]. Exact and approximate expressions for $p_{j,w}$ in (22), for $j = 1,2,3$ and $w = 0,1,2$, have been included in [32] but are also listed in Table 2 for the sake of completeness. They are all functions of the halfway Euclidean distance d' between any two adjacent points along one dimension of the constellation diagram at the receiver.

As explained in Section III, PCH is achieved when N_T is fixed and $N_R \to \infty$, which essentially reduces the mMIMO channel into an equivalent non-fading AWGN channel. In this case, the ergodic SNR is given in (2) and the noise variance per dimension is $\sigma_\mathrm{n}^2/2$. Therefore, d' assumes the form:

$$d' = d\,\frac{\sqrt{\mathrm{snr}}}{\sqrt{0.5}} = d\sqrt{2\,\mathrm{snr}} = \sqrt{\frac{3\,\mathrm{snr}}{M-1}}, \tag{23}$$

where $d = \sqrt{3/(2(M-1))}$ is the halfway distance between adjacent points along one dimension of the M-QAM constellation diagram at the transmitter [46], [47].

Fig. 9 provides an example that demonstrates the memory requirements of symbol-level GRAND. Lookup tables are presented side by side for snr values ranging from 9 dB to 27 dB in steps of 1 dB. Each lookup table contains all possible structures for $w_\mathrm{th} = 4$, arranged in descending order of likelihood as determined by (22). One lookup table is required for each snr value in the range, since the ordering of the error structures depends on snr. In this example, the ordering of the error structures does not change for snr values greater than 18 dB, therefore lookup tables beyond 18 dB can be omitted. Observe in Fig. 9 that $0 \le L_1 \le 4$ and $0 \le L_2 \le 2$, hence L_1 and L_2 can be represented by 3 bits and 2 bits, respectively. A structure $[L_1\ L_2]$ occupies $\lambda = 3+2 = 5$ bits, each lookup table contains $v = 8$ structures, and a total of $\tau = 10$ lookup tables are needed to cover the range between 9 dB and 18 dB in steps of 1 dB. Therefore, symbol-level GRAND will reserve $\lambda v \tau = 400$ bits of memory space in this example.

TABLE 2. Expressions for the probability terms in (22). Function $Q(z) \triangleq (1/\sqrt{2\pi}) \int_z^\infty \exp(-t^2/2)\,dt$ is the tail distribution of the standard normal distribution. Variable d' is given by $d' = \sqrt{3\,\mathrm{snr}/(M-1)}$.

$p_{1,0} = (1-Q(d'))^2$
$p_{2,0} = (1-Q(d'))(1-2Q(d'))$
$p_{3,0} = (1-2Q(d'))^2$
$p_{1,1} = 2(1-Q(d'))Q(d')$
$p_{2,1} \approx [2(1-Q(d'))Q(d') + (1-2Q(d'))Q(d')]$
$p_{3,1} \approx 4(1-2Q(d'))Q(d')$
$p_{1,2} = Q^2(d')$
$p_{2,2} \approx 2Q^2(d')$
$p_{3,2} \approx 4Q^2(d')$

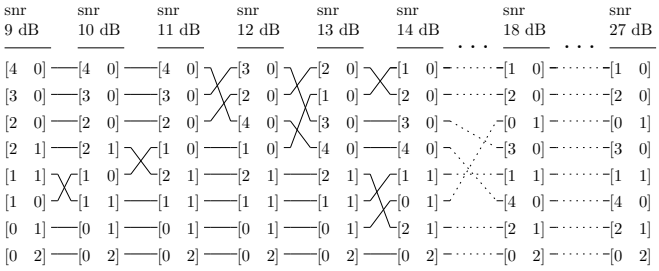

FIGURE 9. Evolution of the ranking of the error structures, based on (22), for $N_R \to \infty$, $w_\mathrm{th} = 4$ and an increasing value of snr.

VII. Results

The performance and the decoding complexity of the considered system were evaluated through numerical simulations. The number of codewords per MIMO channel use, previously defined in Section II, is here set to $N_c = 1$. In this setup, each codeword of n bits is transmitted in "one shot", using Gray-coded M-QAM over a mMIMO channel with $N_T = n/\log_2(M)$ transmit antennas. Both ZF and MMSE detectors have been assessed, using (3) and (4) for filtering and (6) and (7) for antenna sorting. In all considered system configurations, the curves obtained using MMSE detection perfectly matched those obtained using ZF detection, showing no advantage of MMSE over ZF. In the high-SNR regime, both (4) and (7) converge

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3630551

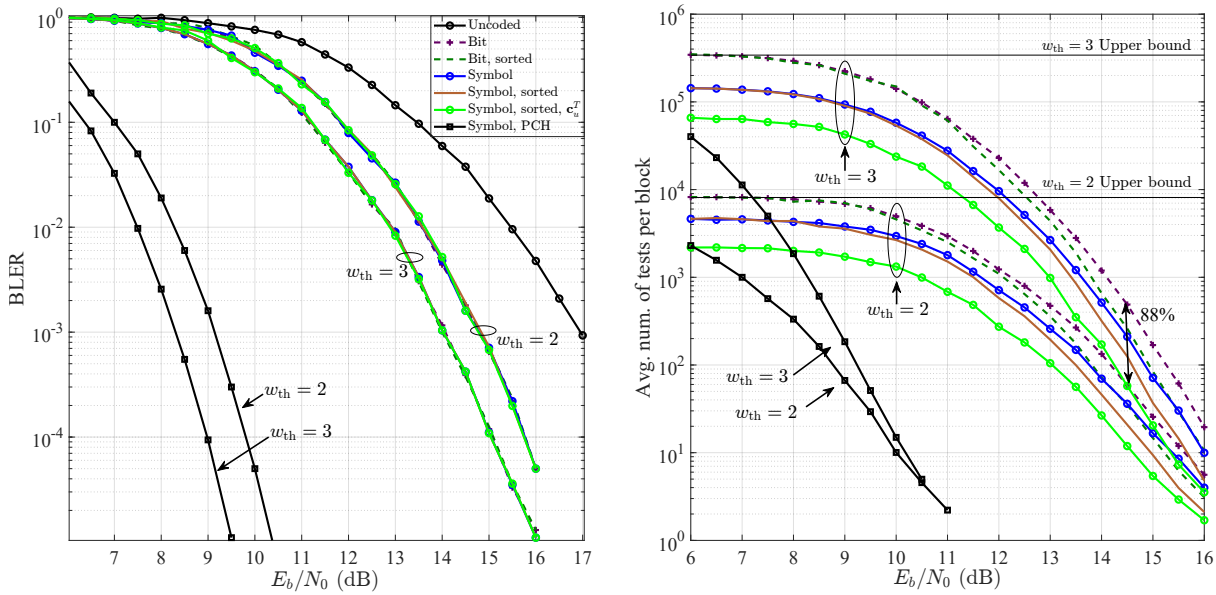

FIGURE 10. **BLER performance (left) and decoding complexity (right) for different thresholds $w_{\text{th}} = 2, 3$, for different decoders, using RLC (128,103), with $N_T = 32$ and $N_R = 50$, and 16-QAM. The corresponding PCH lower bounds are also plotted.**

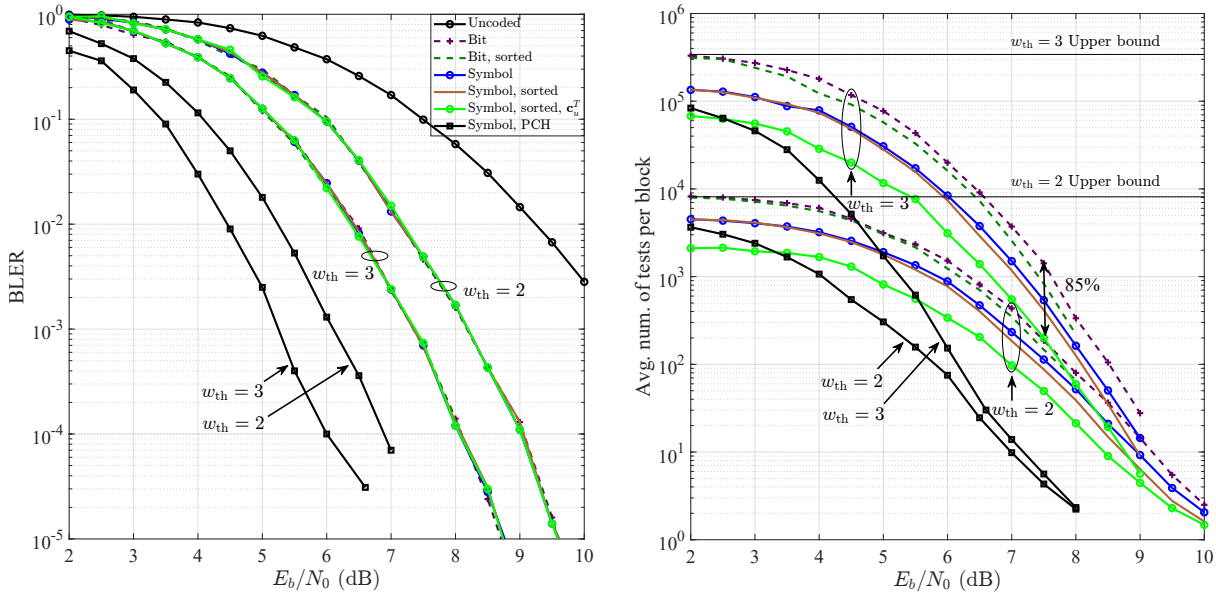

FIGURE 11. **BLER performance (left) and decoding complexity (right) for different thresholds $w_{\text{th}} = 2, 3$, for different decoders, using RLC (128,103), with $N_T = 32$ and $N_R = 100$, and 16-QAM. The corresponding PCH lower bounds are also plotted.**

respectively to (3) and (6). In the low-SNR regime, as seen in Fig. 2, the different filters lead to different antenna sorting orders. However, for a large number of errors in a codeword, the error correction capability of the short codes does not suffice to provide any complexity reduction. Consequently, as MMSE does not offer an advantage over ZF in the considered system model, only results for ZF detection are presented in this section. Note that after the ZF detector, antenna sorting involves the arrangement of the set of g_i, for $i = 1, \cdots, N_T$, in ascending order.

Two constellations were considered: 16-QAM and 64-QAM. As the M-arity of the modulation grows, the number of transmit antennas N_T is decreased to accommodate the same payload of n bits. The number of antennas was set to $N_T = 32$ for 16-QAM with a RLC $(128, 103)$, and $N_T = 22$ for 64-QAM with a RLC $(132, 106)$, such that the code rate $R = k/n = 0.8$ is kept constant. Three different values for N_R were tested for each M-arity: $N_R = 50$, 100 and 200 for 16-QAM, and $N_R = 38$, 69 and 138 for 64-QAM. To make the comparison between 16-QAM and 64-QAM fair, the number of receive antennas N_R in each configuration was chosen to yield similar load factors N_R/N_T, i.e., $50/32 \approx 38/22$, $100/32 \approx 69/22$ and $200/32 \approx 138/22$.

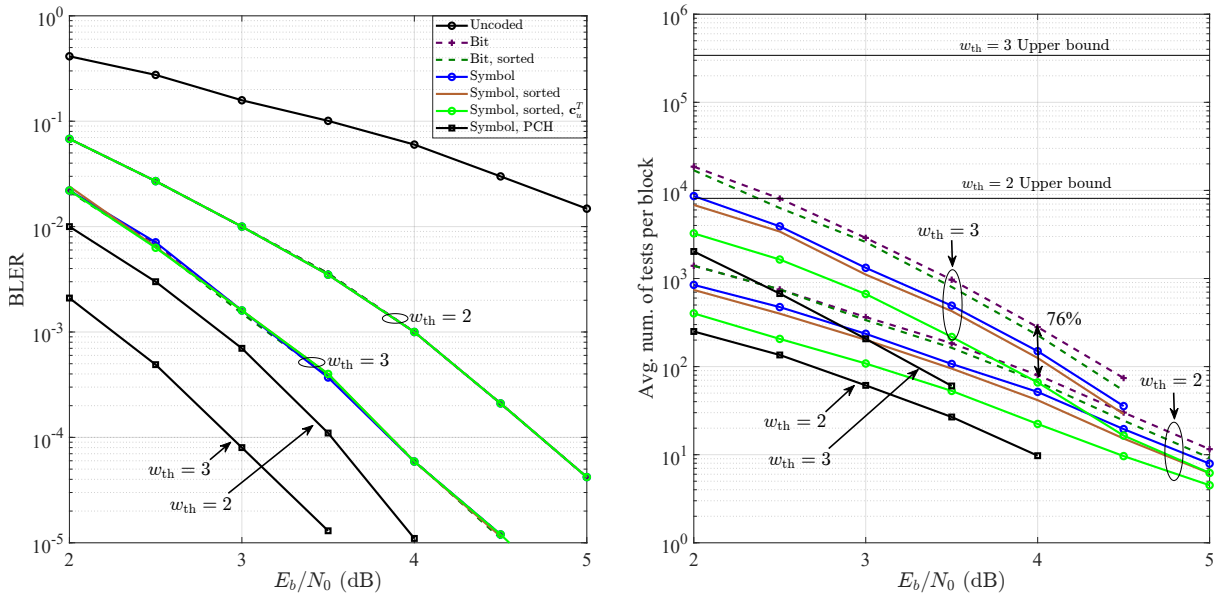

FIGURE 12. **BLER performance (left) and decoding complexity (right) for different thresholds $w_{\rm th} = 2, 3$, for different decoders, using RLC (128,103), with $N_T = 32$ and $N_R = 200$, and 16-QAM. The corresponding PCH lower bounds are also plotted.**

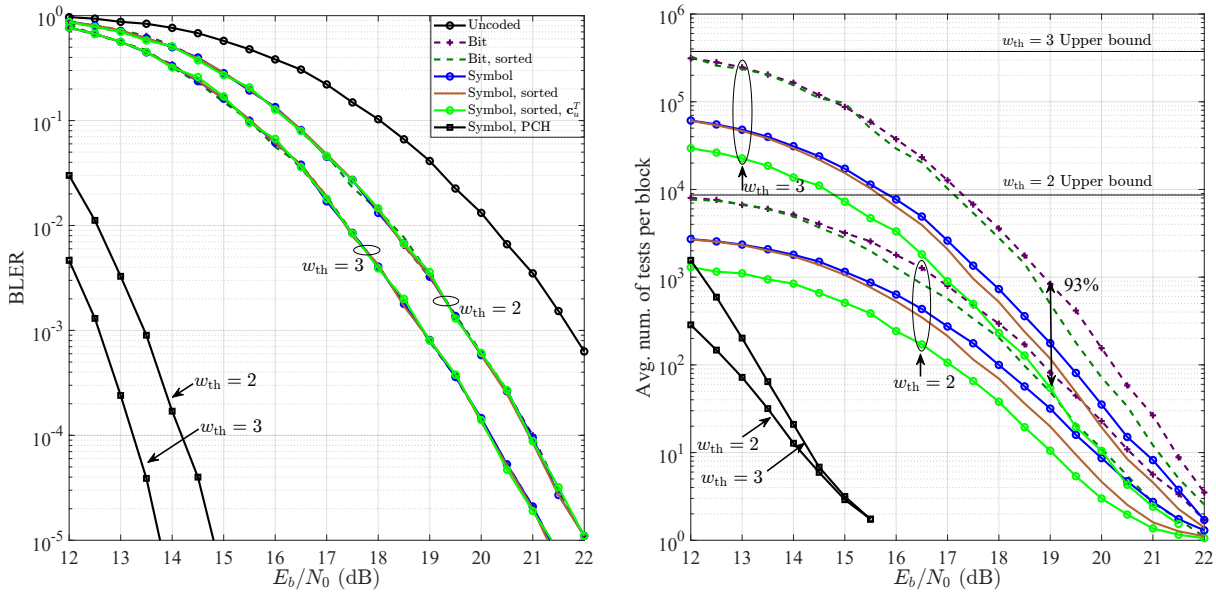

FIGURE 13. **BLER performance (left) and decoding complexity (right) for different thresholds $w_{\rm th} = 2, 3$, for different decoders, using RLC (132,106), with $N_T = 22$ and $N_R = 34$, and 64-QAM. The corresponding PCH lower bounds are also plotted.**

The antenna sorting preprocessing can be used to improve the decoding speed of symbol-level GRAND but also of bit-level GRAND. After arranging the bit-stings $\mathcal{S}(\tilde{s}_i)$ in ascending order of likelihood, one can also apply the original bit-level GRAND using its default flipping order for each bit. However, this leads to sub-optimal performance, given that the probability of the strings of $\log_2(M)$ bits is being used rather than the probability of the individual bits. We refer to this decoding method as *sorted-bit-level decoding*. While sub-optimal, this ordering performs a step toward optimal bit ordering, and therefore reduces decoding complexity in comparison to standard unsorted bit-level GRAND. Afterward, decoding can be further accelerated by using the row of the parity-check matrix conveying most information about the errors' location as a filter on the candidate error patterns coming from the previous decoding steps as described in Section V.

For clarity and completeness, Table 3 summarizes all the labels that appear in the legends of the simulation figures. Fig. 10, Fig. 11 and Fig. 12 illustrate the performance and decoding complexity results for 16-QAM, and Fig. 13, Fig. 14 and Fig. 15 show the performance and decoding complexity results for 64-QAM. The block error rate (BLER) has been used to evaluate the system's performance as a function of

This article has been accepted for publication in IEEE Open Journal of the Communications Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCOMS.2025.3630551

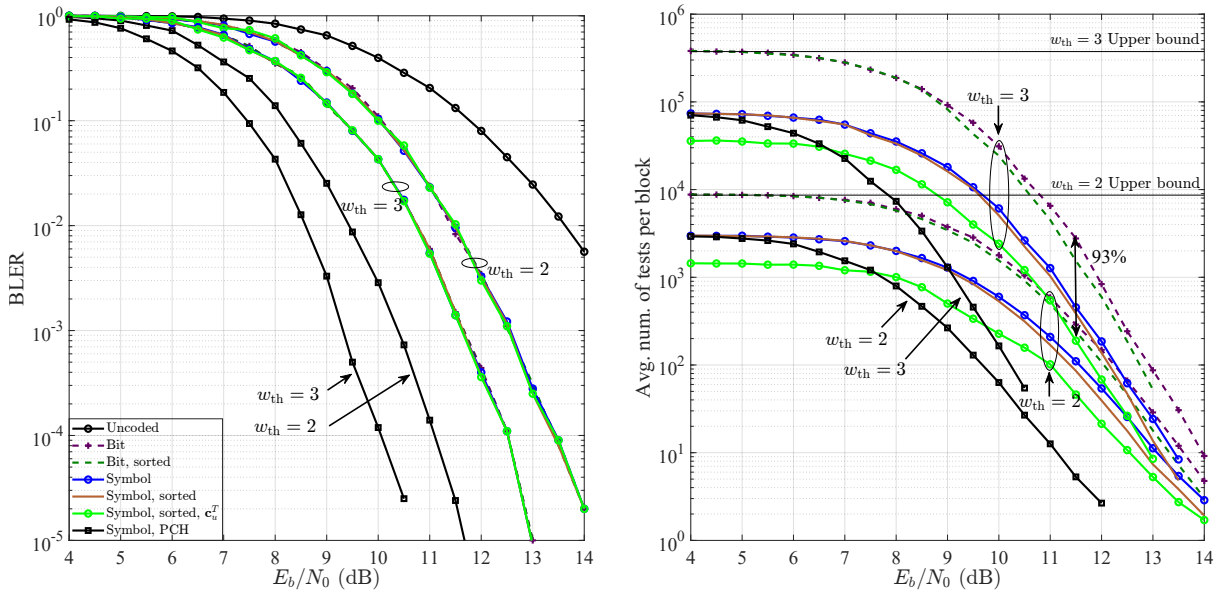

FIGURE 14. **BLER performance (left) and decoding complexity (right) for different thresholds $w_{\rm th} = 2, 3$, for different decoders, using RLC (132,106), with $N_T = 22$ and $N_R = 69$, and 64-QAM. The corresponding PCH lower bounds are also plotted.**

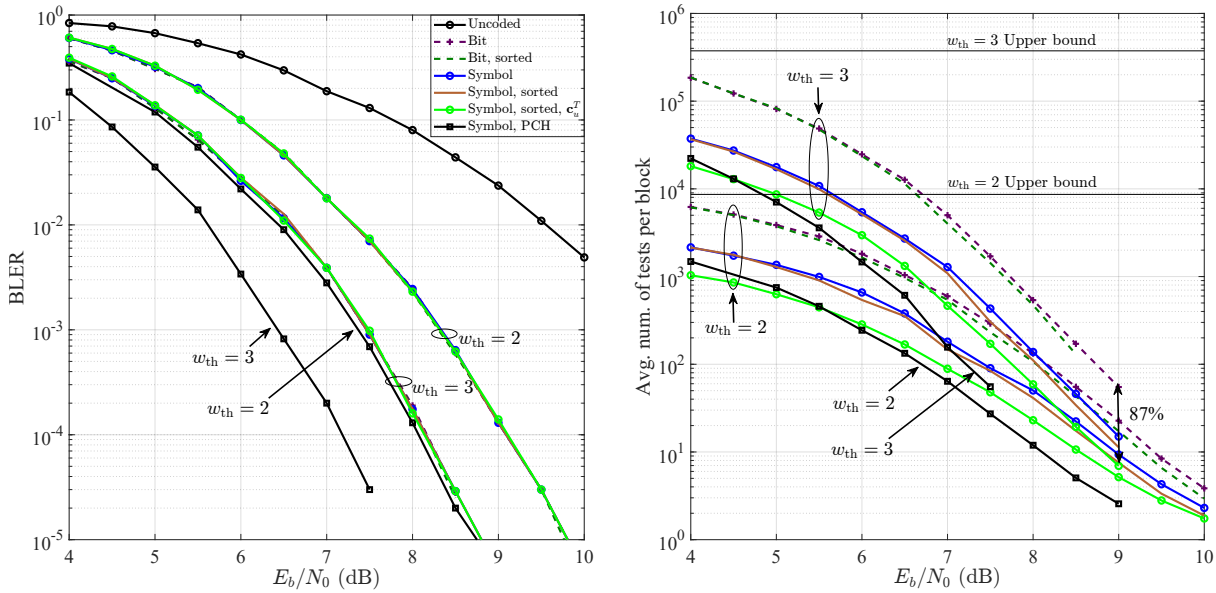

FIGURE 15. **BLER performance (left) and decoding complexity (right) for different thresholds $w_{\rm th} = 2, 3$, for different decoders, using RLC (132,106), with $N_T = 22$ and $N_R = 138$, and 64-QAM. The corresponding PCH lower bounds are also plotted.**

E_b/N_0, as it is commonly used in recent works evaluating GRAND. The decoding complexity has been expressed in terms of the expected number of membership tests needed at each E_b/N_0. All figures include the curves for uncoded transmission, bit-level GRAND decoding, sorted-bit-level decoding, symbol-level GRAND decoding, sorted-symbol-level decoding, and sorted-symbol-level decoding with extra "filtering" provided by the application of the lightweight membership test. Each system configuration is assessed with two different thresholds for the number of bits in error in the error pattern, $w_{\rm th} = 2, 3$. The figures also include performance and complexity results when using symbol-level GRAND for PCH. Recall that, in the ideal scenario of PCH, antenna sorting prior to GRAND has no impact on the overall decoding complexity because all streams experience the same SNR after ZF detection, as seen in (14).

As expected, the BLER performance greatly improves as one tests error patterns with larger weights, but this is achieved at the cost of a considerably larger number of membership tests. When considering a given $w_{\rm th}$ threshold, the worst-case upper bound on the number of membership tests occurs in the case of bit-level GRAND, which is $\sum_{w=0}^{w_{\rm th}} \binom{n}{w}$. The upper-bound value for $w_{\rm th} = 2, 3$ has been plotted in the figures that show complexity results. The

TABLE 3. Summary of the simulation curves presented in the figures.

Legend	Description
Uncoded	Baseline case without channel coding. Performance reflects only MIMO detection.
Bit	Conventional bit-level GRAND decoding combined with ZF detection. The test error patterns are generated at the bit level.
Bit, sorted	Bit-level GRAND decoding assisted by antenna sorting.
Symbol	Symbol-level GRAND decoding (no sorting). The test error patterns are generated at the symbol level (i.e. strings of length $\log_2(M)$).
Symbol, sorted	Symbol-level GRAND with test error patterns guided by the antenna sorting order.
Symbol, sorted, $\mathbf{c}_u^T$	Constrained symbol-level GRAND (sorted) with the proposed lightweight membership test based on the optimized row $\mathbf{c}_u^T$.
Symbol, PCH	Lower bound under perfect channel hardening (PCH) assumption, where all spatial streams are equally reliable and antenna sorting brings no benefit.

results show that the average number of membership tests is much lower than the upper bound for $M = 16$ with $N_R = 200$ and $M = 64$ with $N_R = 138$. Nevertheless, when the noise is too large, the decoding complexity can get close to the upper bound due to the sheer number of erroneous symbols. As one would expect, when the noise vanishes, the average number of membership tests always tends to be one; in that case, all the received words are valid codewords and the only membership test performed is to check, and confirm that the error pattern is $\hat{\mathbf{e}}_\mathrm{b} = \mathbf{0}$.

One should note that the BLER performance results for bit-level GRAND, symbol-level GRAND, sorted-bit-level, sorted-symbol-level, and sorted-constrained-symbol-level decoding for the analyzed range of E_b/N_0 are all the same (all four curves overlap). However, the complexity comparison illustrates that the sorting antenna schemes and then symbol-level sorted with the lightweight test involving $\mathbf{h}_u^T$ remarkably outperform bit-level GRAND. The extra complexity reduction added by the sorting preprocessing becomes more significant when N_R becomes smaller. This is due to a less strong channel hardening effect so that the equivalent SNR at each of the N_T streams becomes more uneven. As seen in the complexity results, the decoding time is reduced by $\approx 80\%$ when $M = 16$, and by $\approx 90\%$ in the system with 64-QAM.

VIII. Conclusion

This paper proposes a coded mMIMO transmission scheme for high-throughput, high-reliability, and low-latency, in accordance to the URLLC desiderata. This is accomplished using RLCs and ordered reliability symbol-level GRAND. Symbol-level GRAND has been generalized from M-QAM to any M-ary modulation scheme, and extended from SISO channels to mMIMO channels.

When the channel conditions are far from ideal, the paper shows that linear detectors can provide a soft-metric for the reliability of each spatial stream, which in the proposed setup corresponds to a symbol reliability. The orthogonality defect of the mMIMO lattice is related to the variance of the reliability of the symbols. The disparity between the reliability of the symbols gets larger when N_R decreases; in that case, the proposed antenna sorting can provide a significant reduction of the decoding complexity.

The results show that symbol-level GRAND provides much faster decoding than the bit-level GRAND counterpart in the same mMIMO setup, throughout the SNR range of interest. The proposed antenna sorting mechanism further speeds up the decoding process in scenarios where channel-hardening is far from perfect. Furthermore, the error patterns generated by the symbol-level decoder with antenna sorting decoding stages undergo a "filtering" constraint based on a partial membership test. This proposed lightweight membership test employs a linear complexity constraint $\mathcal{O}(n)$, rather than the quadratic one, $\mathcal{O}(n^2)$, involved in a full membership test. This technique has been optimized to maximize the number of discarded error patterns. In the worst scenario of "filtering" with a balanced row, half of the error patterns are discarded.

Without performance degradation, the number of membership tests can be dramatically reduced when comparing the results of the GRAND receiver combining symbol-level GRAND, antenna sorting and the lightweight membership test, with the results of using conventional bit-level GRAND.

Appendix: Antenna SNR with perfect and imperfect CSI

This appendix explores the impact of imperfect CSI on the output SNR at each antenna stream for ZF. For a clearer and more streamlined presentation, the derivation of (6) for perfect CSI is first presented and a similar line of thought is then followed to obtain an expression for the output SNR in the case of imperfect CSI.

A. Perfect CSI

While the case with perfect CSI is well-known (e.g., [33]), we start by deriving (6) with the aim of obtaining useful equalities for the following subsection. Let us consider the model in (5). After applying (3) (and dropping the ZF notation), one gets $\mathbf{W}\mathbf{y} = \mathbf{x} + \mathbf{W}\mathbf{n}$. The post-detection noise covariance matrix is

$$\begin{aligned}\mathbf{R}_n &= \mathbb{E}\left[\mathbf{W}\mathbf{n}\mathbf{n}^H\mathbf{W}^H\right] \\ &= \mathbf{W}\underbrace{\mathbb{E}\left[\mathbf{n}\mathbf{n}^H\right]}_{=\sigma_\mathrm{n}^2\mathbf{I}}\mathbf{W}^H = \sigma_\mathrm{n}^2\mathbf{W}\mathbf{W}^H. \end{aligned} \tag{24}$$

It is possible to show that, using (3),

$$\mathbf{W}\mathbf{W}^H = \left(\mathbf{H}^H\mathbf{H}\right)^{-1}, \tag{25}$$

and therefore, considering $\mathbb{E}\{|s_i|^2\} = 1$, the resulting SNR for the i-th stream is

$$\mathrm{snr}_i^{(\text{perfect CSI})} = \frac{1}{\sigma_{\mathrm{n}}^2 \left[\left(\mathbf{H}^H\mathbf{H}\right)^{-1}\right]_{ii}} = \frac{\mathrm{snr}}{[\mathbf{G}^{-1}]_{ii}} = g_i \ \mathrm{snr}. \tag{26}$$

B. Imperfect CSI

In the case of imperfect CSI, the estimated channel matrix can be modeled as

$$\widehat{\mathbf{H}} = \mathbf{H} + \mathbf{H}_\varepsilon, \tag{27}$$

where the elements of $\mathbf{H}_\varepsilon$ are i.i.d. $\mathcal{CN}(0, \sigma_\varepsilon^2)$. The received signal can be written as

$$\mathbf{y} = \left(\widehat{\mathbf{H}} - \mathbf{H}_\varepsilon\right)\mathbf{x} + \mathbf{n}. \tag{28}$$

The detection filter constructed at the receiver will then be

$$\widehat{\mathbf{W}} = \left(\widehat{\mathbf{H}}^H\widehat{\mathbf{H}}\right)^{-1}\widehat{\mathbf{H}}^H. \tag{29}$$

Hence, the detected signal is

$$\begin{aligned}\widehat{\mathbf{W}}\mathbf{y} &= \left(\widehat{\mathbf{H}}^H\widehat{\mathbf{H}}\right)^{-1}\hat{\mathbf{H}}^H\left(\widehat{\mathbf{H}} - \mathbf{H}_\varepsilon\right)\mathbf{x} + \widehat{\mathbf{W}}\mathbf{n} \\ &= \mathbf{x} - \widehat{\mathbf{W}}\mathbf{H}_\varepsilon\mathbf{x} + \widehat{\mathbf{W}}\mathbf{n}.\end{aligned} \tag{30}$$

One has now to compute the power of the two noise terms, starting by obtaining the correlation matrices

$$\begin{aligned}\mathbf{R}_\varepsilon &= \mathbb{E}\left[\widehat{\mathbf{W}}\mathbf{H}_\varepsilon\mathbf{H}_\varepsilon^H\widehat{\mathbf{W}}^H\right] \\ &= \widehat{\mathbf{W}}\underbrace{\mathbb{E}\left[\mathbf{H}_\varepsilon\mathbf{H}_\varepsilon^H\right]}_{=\sigma_\varepsilon^2\mathbf{I}}\widehat{\mathbf{W}}^H \\ &= \sigma_\varepsilon^2\widehat{\mathbf{W}}\widehat{\mathbf{W}}^H = \sigma_\varepsilon^2\left(\widehat{\mathbf{H}}^H\widehat{\mathbf{H}}\right)^{-1},\end{aligned} \tag{31}$$

where the last equality uses (25). The covariance matrix of the $\widehat{\mathbf{W}}\mathbf{n}$ noise term is obtained similarly to (24) and (25)

$$\mathbf{R}_{\hat{n}} = \mathbb{E}\left[\widehat{\mathbf{W}}\mathbf{n}\mathbf{n}^H\widehat{\mathbf{W}}^H\right] \tag{32}$$

$$= \sigma_{\mathrm{n}}^2\widehat{\mathbf{W}}\widehat{\mathbf{W}}^H = \sigma_{\mathrm{n}}^2\left(\widehat{\mathbf{H}}^H\widehat{\mathbf{H}}\right)^{-1}. \tag{33}$$

Therefore, the SNR for the i-th stream under imperfect CSI is

$$\begin{aligned}\mathrm{snr}_i^{(\text{imperfect CSI})} &= \frac{1}{\left[\left(\widehat{\mathbf{H}}^H\widehat{\mathbf{H}}\right)^{-1}(\sigma_\varepsilon^2 + \sigma_{\mathrm{n}}^2)\right]_{ii}} \\ &= \frac{\frac{1}{\sigma_{\mathrm{n}}^2}}{\hat{g}_i^{-1}\left(1 + \frac{\sigma_\varepsilon^2}{\sigma_{\mathrm{n}}^2}\right)} \\ &= \frac{\mathrm{snr}}{\hat{g}_i^{-1}\left(1 + \sigma_\varepsilon^2\mathrm{snr}\right)},\end{aligned} \tag{34}$$

where $\mathrm{snr} = \frac{1}{\sigma_{\mathrm{n}}^2}$. Also, $\hat{g}_i = [\widehat{\mathbf{G}}]_{ii}$, and $\widehat{\mathbf{G}} = \widehat{\mathbf{H}}^H\widehat{\mathbf{H}}$ stands for the imperfectly estimated Gram matrix.

Note that when $\sigma_\varepsilon^2 = 0$, one recovers (6), as in that case $\hat{g}_i = g_i$. The impact of a small σ_ε^2 on the ordering of the antennas will be negligible, however for large perturbations of the channel estimate, with $\sigma_\varepsilon^2 \gg \sigma_{\mathrm{n}}^2$, the estimates in (34) are defined by the random variables $\hat{g}_i$. While under perfect CSI, the sorting order is deterministically obtained from $\mathbf{H}$, when only a very poor estimate of the channel is available, the sorting becomes random. For near-perfect CSI, the sorting order should be nearly identical or even equal to the optimal order. The evaluation of how the decoding performance is affected and how robust the complexity reduction remains for a given level of CSI uncertainty is suggested for future work.

REFERENCES

[1] P. Nouri, H. Alves, M. A. Uusitalo, O. Alcaraz López, and M. Latva-aho, "Machine-type wireless communications enablers for beyond 5G: Enabling URLLC via diversity under hard deadlines," *Computer Networks*, vol. 174, no. 3, p. 107227, Jun. 2020.

[2] W. An, M. Médard, and K. R. Duffy, "Keep the bursts and ditch the interleavers," *IEEE Transactions on Communications*, vol. 70, no. 6, pp. 3655–3667, May 2022.

[3] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, Jul. 1948.

[4] R. Gallager, "The random coding bound is tight for the average code (Corresp.)," *IEEE Transactions on Information Theory*, vol. 19, no. 2, pp. 244–246, Mar. 1973.

[5] M. Shirvanimoghaddam, M. S. Mohammadi, R. Abbas, A. Minja, C. Yue, B. Matuz, G. Han, Z. Lin, W. Liu, Y. Li, S. Johnson, and B. Vucetic, "Short block-length codes for ultra-reliable low latency communications," *IEEE Communications Magazine*, vol. 57, no. 2, pp. 130–137, Feb. 2019.

[6] A. Becker, A. Joux, A. May, and A. Meurer, "Decoding random binary linear codes in $2^{n/20}$: How $1 + 1 = 0$ improves information set decoding," in *Proc. 31st Int. Conf. Theory and App. of Crypt. Techn. (EUROCRYPT)*, Cambridge, United Kingdom, Apr. 2012.

[7] J. Wolf, "Efficient maximum likelihood decoding of linear block codes using a trellis," *IEEE Transactions on Information Theory*, vol. 24, no. 1, pp. 76–80, Jan. 1978.

[8] T. Kasami, T. Takata, T. Fujiwara, and S. Lin, "On complexity of trellis structure of linear block codes," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 1057–1064, May 1993.

[9] F. R. Kschischang and V. Sorokine, "On the trellis structure of block codes," *IEEE Transactions on Information Theory*, vol. 41, no. 6, pp. 1924–1937, Nov. 1995.

[10] G. D. Forney, "Coset codes - Part II: Binary lattices and related codes," *IEEE Transactions on Information Theory*, vol. 34, no. 5, pp. 1152–1187, Sep. 1988.

[11] A. H. Banihashemi, "Decoding Complexity and Trellis Structure of Lattices," Ph.D. dissertation, University of Waterloo, 1997.

[12] F. A. Monteiro and F. R. Kschischang, "Trellis detection for random lattices," in *Proc. of 8th Inter. Symp. on Wireless Communication Systems (ISWCS)*, Aachen, Germany, Nov. 2011, pp. 755–759.

[13] B. Honary, *Trellis Decoding of Block Codes: A Practical Approach*. New York, USA: Kluwer Academic Publishers, 1997.

[14] S. Lin, T. Kasami, and M. Fossorier, *Trellises and Trellis-Based Decoding Algorithms for Linear Block Codes*. New York, USA: Kluwer Academic Publishers, 1998.

[15] J. T. Coffey and R. M. Goodman, "The complexity of information set decoding," *IEEE Transactions on Information Theory*, vol. 36, no. 5, pp. 1031–1037, Sep. 1990.

[16] K. R. Duffy, J. Li, and M. Médard, "Capacity-achieving guessing random additive noise decoding," *IEEE Transactions on Information Theory*, vol. 65, no. 7, pp. 4023–4040, Jul. 2019.

[17] K. R. Duffy, A. Solomon, K. M. Konwar, and M. Médard, "5G NR CA-polar maximum likelihood decoding by GRAND," *54th Annual Conf. on Info. Sciences and Systems (CISS)*, May 2020.

[18] K. R. Duffy, "Ordered reliability bits guessing random additive noise decoding," in *Proc. of IEEE Inter. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, Jun 2021, p. 8268–8272.

[19] S. M. Abbas, M. Jalaleddine, and W. J. Gross, "GRAND for Rayleigh fading channels," in *2022 IEEE Globecom Workshops (GC Wkshps)*, Rio de Janeiro, Brazil, Jan. 2022, pp. 504–509.

[20] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, Apr. 2010.

[21] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4232–4265, Apr. 2014.

[22] C. Yue, V. Miloslavskaya, M. Shirvanimoghaddam, B. Vucetic, and Y. Li, "Efficient decoders for short block length codes in 6G URLLC," *IEEE Communications Magazine*, vol. 61, no. 4, pp. 84–90, Apr. 2023.

[23] S. Allahkaram, F. A. Monteiro, and I. Chatzigeorgiou, "URLLC with Coded Massive MIMO via Random Linear Codes and GRAND," in *IEEE 96th Vehicular Technology Conference (VTC-Fall)*, London, UK, Sep. 2022, pp. 1–5.

[24] K. R. Duffy, M. Médard, and W. An, "Guessing random additive noise decoding with symbol reliability information (SRGRAND)," *IEEE Transactions on Communications*, vol. 70, no. 1, pp. 3–18, Sep. 2022.

[25] D. Cruz, F. A. Monteiro, and B. C. Coutinho, "Quantum error correction via noise guessing decoding," *IEEE Access*, vol. 11, pp. 19 446 – 119 461, Oct. 2023.

[26] D. Cruz, F. A. Monteiro, A. Roque, and B. C. Coutinho, "Fault-tolerant noise guessing decoding of quantum random codes," *IEEE Transactions on Quantum Engineering*, vol. 6, pp. 2689–1808, Aug. 2025.

[27] D. Chandra, Z. B. Kaykac Egilmez, Y. Xiong, S. X. Ng, R. G. Maunder, and L. Hanzo, "Universal decoding of quantum stabilizer codes via classical guesswork," *IEEE Access*, vol. 11, pp. 19 059–19 072, 2023.

[28] X. Ma, "Guessing what, noise or codeword?" in *2024 IEEE Information Theory Workshop (ITW)*, Shenzhen, China, Nov. 2024, pp. 460–465.

[29] R. Hadavian, D. Truhachev, K. El-Sankary, H. Ebrahimzad, and H. Najafi, "Ordered reliability direct error pattern testing (ORDEPT) algorithm," in *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, Kuala Lumpur, Malaysia, Dec. 2023, pp. 6983–6988.

[30] A. Solomon, K. R. Duffy, and M. Médard, "Soft Maximum Likelihood Decoding using GRAND," in *Proc. of IEEE International Conference on Communications (ICC)*, virtual conf., Jun. 2020.

[31] P. Yuan, M. Médard, K. Galligan, and K. R. Duffy, "Soft-output (so) grand and iterative decoding to outperform ldpc codes," *IEEE Transactions on Wireless Communications*, vol. 24, no. 4, pp. 3386–3399, Jan. 2025.

[32] I. Chatzigeorgiou and F. A. Monteiro, "Symbol-Level GRAND for High-Order Modulation Over Block Fading Channels," *IEEE Communications Letters*, vol. 27, no. 2, pp. 447–451, Feb. 2023.

[33] Y. Jiang, M. K. Varanasi, and J. Li, "Performance analysis of ZF and MMSE equalizers for MIMO systems: An in-depth study of the high SNR regime," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2008–2026, Mar. 2011.

[34] F. A. Monteiro, "Lattices in MIMO spatial multiplexing: Detection and geometry," Ph.D. dissertation, University of Cambridge, United Kingdom, 2012.

[35] K. Su and I. Wassell, "A new ordering for efficient sphere decoding," in *IEEE International Conference on Communications (ICC)*, vol. 3, May 2005, pp. 1906–1910 Vol. 3.

[36] J. Griffin, P. Yuan, K. R. Duffy, and M. Medard, "Using a single-parity-check to reduce the guesswork of guessing codeword decoding," *arXiv:2411.09803 [cs.IT]*, Nov. 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2411.09803

[37] M. Rowshan and J. Yuan, "Constrained error pattern generation for GRAND," in *2022 IEEE International Symposium on Information Theory (ISIT)*, Espoo, Finland, Jun. 2022, pp. 1767–1772.

[38] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge, UK: Cambridge University Press, 2003.

[39] R. W. Heath Jr. and A. Lozano, *Foundations of MIMO Communication*. Cambridge, UK: Cambridge University Press, 2019.

[40] F. A. Monteiro and I. J. Wassell, "Recovery of a lattice generator matrix from its Gram matrix for feedback and precoding in MIMO," in *4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, Limassol, Cyprus, 2010, pp. 1–6.

[41] M. Taherzadeh, A. Mobasher, and A. K. Khandani, "LLL reduction achieves the receive diversity in MIMO decoding," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4801–4805, Dec 2007.

[42] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Foundations and Trends® in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.

[43] S.-J. Park, "Triangular quadrature amplitude modulation," *IEEE Commun. Lett.*, vol. 11, no. 4, pp. 292–294, Apr. 2007.

[44] S.-J. Park and M.-K. Byeon, "Irregularly distributed triangular quadrature amplitude modulation," in *Proc. IEEE 19th Int. Symp. Pers., Indoor Mobile Radio Commun.*, Cannes, France, 2008, pp. 1–5.

[45] T. K. Oikonomou, S. A. Tegos, D. Tyrovolas, P. D. Diamantoulakis, and G. K. Karagiannidis, "On the error analysis of hexagonal-QAM constellations," *IEEE Commun. Lett.*, vol. 26, no. 8, pp. 1764–1768, Aug. 2022.

[46] J. Lu, K. B. Letaief, J.-I. Chuang, and M. L. Liou, "M-PSK and M-QAM BER computation using signal-space concepts," *IEEE Transactions on communications*, vol. 47, no. 2, pp. 181–184, Feb 1999.

[47] K. Cho and D. Yoon, "On the general BER expression of one- and two-dimensional amplitude modulations," *IEEE Transactions on Communications*, vol. 50, no. 7, pp. 1074–1080, Jul. 2002.

SAHAR ALLAHKARAM is a PhD student in the Dep. of Information Science and Technology at Iscte – University Institute of Lisbon, Portugal. She is working on Signal Processing and Coding Techniques for 6G Ultra-reliable Low-latency Wireless Machine-type Communications. She obtained her MSc in Aerospace Engineering from Sapienza University of Rome in 2020 and her BSc in Electronic Engineering from Azad University of Tehran in 2015.

FRANCISCO A. MONTEIRO (Member, IEEE) is Associate Professor in the Dep. of Information Science and Technology at Iscte – University Institute of Lisbon, and a researcher at Instituto de Telecomunicações, Lisbon, Portugal. He holds a PhD from the University of Cambridge, UK, and the Licenciatura and MSc degrees in Electrical and Computer Engineering from Instituto Superior Técnico (IST), University of Lisbon, where he also became a Teaching Assistant. He held visiting research positions at the Universities of Toronto (Canada), Lancaster (UK), Oulu (Finland), and Pompeu Fabra (Barcelona, Spain). He has won two best paper prizes awards at IEEE conferences (2004 and 2007), a Young Engineer Prize (3rd place) from the Portuguese Engineers Institution (Ordem dos Engenheiros) in 2002, and for two years in a row was a recipient of Exemplary Reviewer Awards from the IEEE Wireless Communications Letters (in 2014 and in 2015). He co-edited the book "MIMO Processing for 4G and Beyond: Fundamentals and Evolution", published by CRC Press in 2014. In 2016 he was the Lead Guest Editor of a special issue on Network Coding of the EURASIP Journal on Advances in Signal Processing. He was a general chair of ISWCS 2018 - The 15th International Symposium on Wireless Communication Systems, an IEEE major conference in wireless communications.

IOANNIS CHATZIGEORGIOU (Senior Member, IEEE) received the Dipl.Ing. degree in Electrical Engineering from Democritus University of Thrace, Greece, the M.Sc. degree in Satellite Communication Engineering from the University of Surrey, UK, and the Ph.D. degree from the University of Cambridge, UK. He is currently a Senior Lecturer at Lancaster University, UK. Prior to his appointment, he held postdoctoral positions at the University of Cambridge and the Norwegian University of Science and Technology, supported by the Engineering and Physical Sciences Research Council (EPSRC) and the European Research Consortium for Informatics and Mathematics (ERCIM), respectively. His research interests include signal processing, coding theory and performance analysis of communication systems.